UK Speech Conference

Sheffield

20–21 June 2016

The University Of Sheffield.

Programme

The technical programme contains 3 keynotes, 1 oral session with 5 presentations and 3 poster sessions with 62 total posters.

Monday June 20th

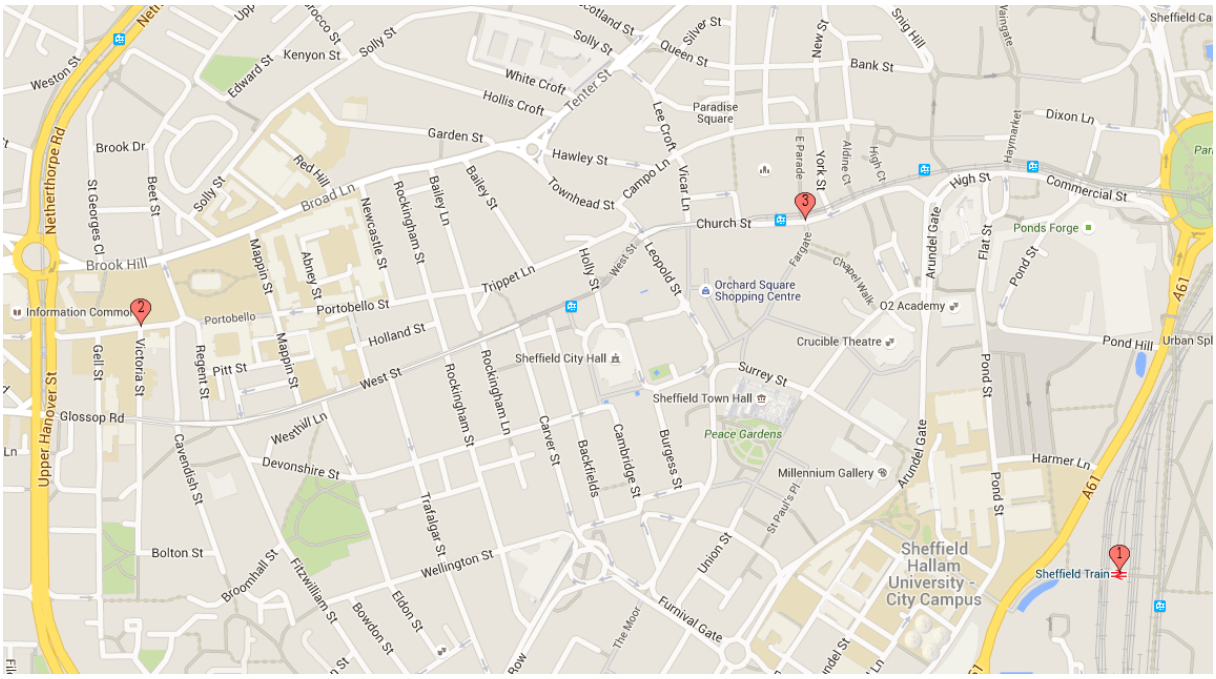
Time	Location	Session
12:00-13:30	Diamond - Workroom 2	Registration and Lunch
13:30-13:45	Diamond - Lecture Theatre 1	Welcoming and introduction to the conference
13:45-14:45	Diamond - Lecture Theatre 1	Keynote: *Computer Recognition of Children's Speech* – Martin Russell, University of Birmingham
14:45-16:00	Diamond - Workroom 1	Poster Session 1
16:00-16:20	Diamond - Workroom 2	Break
16:20-18:00	Diamond - Lecture Theatre 1	Oral Session
19:00-19:30	Cutler's Hall	Drinks reception
19:30-	Cutler's Hall	Buffet dinner

Tuesday June 21st

Time	Location	Session
9:00-10:00	Diamond - Lecture Theatre 1	Keynote: *Speech Perception and the Evaluation of Synthetic Speech* – Mirjam Wester, University of Edinburgh
10:00-11:15	Diamond - Workroom 1	Poster Session 2
11:15-11:45	Diamond - Ground Floor	Break
11:45-13:00	Diamond - Workroom 1	Poster Session 3
13:00-14:00	Diamond - Ground Floor	Lunch
14:00-15:00	Diamond - Lecture Theatre 1	Keynote: *Clinical Challenges for Speech Technology* – Phil Green, University of Sheffield
15:00-15:15	Diamond - Lecture Theatre 1	Final remarks and farewell

Map and Directions

The following is a map of Sheffield city centre. It shows the two points where the conference activities will be located (The Diamond and Cutler's Hall) as well as the location of Sheffield train station.



1. Sheffield train station (Sheaf St, S1 2BP)
2. The Diamond (32 Leavygreave Rd, S3 7RD)
   Lecture Theatre 1 is located in the basement of the building
   Workrooms 1 & 2 are located in the ground floor of the building
3. Cutler's Hall Hospitality (Church St, S1 1HG)

The Venue

UKSpeech 2016 is hosted in the Diamond, opened on September 2015 and the largest ever investment of the University of Sheffield in teaching and learning. It won the Design through Innovation award in the 2016 Yorkshire and Humber Region Royal Institute of Chartered Surveyors (RICS) awards, and was also shortlisted for the Yorkshire awards from the Royal Institute of British Architects (RIBA). The Diamond has an aluminium diamond shaped facade exterior, with galvanised steel sheets and glass; all of which have been recycled. It has been designed as a "smart" building allowing detailed control of energy management, and includes a central naturally ventilated atrium and rainwater harvesting.

Inside, the six-storey Diamond boasts specialist teaching facilities including a range of lecture theatres, seminar rooms, open-plan learning spaces, library and IT services, and space for informal study including a cafe. The computing area offers 1,000 study spaces available 24/7 for all students and staff across the University. There are also digital and print facilities, media editing booths, a recording studio and computer teaching laboratories. The buildings 19 laboratories offer students more practical learning opportunities with a chemical engineering pilot plant, a clean room, an aerospace simulation lab and a virtual reality suite.

Social Events

On Monday 20th, there will a sponsored drinks reception and sit–down buffet dinner at Sheffield's Cutlers' Hall. The present building is the third hall employed by the Company of Cutlers in Hallamshire and was built in 1832 by Samuel Worth and Benjamin Broomhead Taylor. It is currently a Grade II* listed building and it is regarded as one of the finest livery halls in the north of England. The building lies on the same site of the first Hall, which was built in 1638, following the incoporation of the Company of Cutlers in 1621.

From its beginnings in the early part of the 17th century, the Cutlers' Company has placed an integral role both in the expansion of the major local industries and in the growth of Sheffield. The Company of Cutlers consists of an annually elected group of thirty-three people a Master Cutler, two Wardens, six Searchers and twenty-four Assistants. A Clerk and a Beadle are employed for administration and to perform ceremonial duties. The first Company came into existence in August, 1624 with Robert Sorby as its first Master Cutler. In the following August, the Company met to elect the next Company, checked the Master's accounts, installed the new Master and heard a service at the parish church. A new Company has been elected every year to the present day, except during the First and Second World Wars when the same Company continued for the duration of the wars.

Keynotes

Computer Recognition of Children's Speech
Martin Russell, University of Birminigham, Monday June 20th, 13:45

It is has been known since the mid–1990s that automatic speech recognition (ASR) is more challenging for child speech compared with adult speech. Spectral structures, such as formants, occur at higher frequencies in childrens speech, due to children's shorter vocal tracts, and spectral resolution is poor due to their higher fundamental frequencies. In addition, a number of studies have demonstrated greater variability in a range of acoustic parameters in children's speech. However, whether this variability is due to poor motor control or cognitive, phonological factors associated with language acquisition is not known at present. Whatever the cause of this variability, ASR error rates are typically more than 100% greater for child speech compared with adult speech on similar tasks. This is unfortunate, because ASR has a number of compelling applications with children. Moreover, for some of these applications, such as pronunciation or reading tuition, ASR is the key enabling technology, and not just an alternative means of interaction. A number of these applications are particularly demanding because they require accurate recognition at the phone level. This talk will chart progress in ASR for childrens speech from the early work in the mid-1990s to the most recent DNN-HMM based systems. Rather than focussing only on improvements in recognition accuracy, I will also try to measure progress in terms of how our understanding of variability in children's speech has improved, and how research in speech technology might even give new insights into the nature of childrens speech. I will finish by outlining what I think are the most interesting challenges for the future and how they might be addressed.

Speech Perception and the Evaluation of Synthetic Speech
Mirjam Wester, University of Edinburgh,Tuesday June 21st, 9:00

In this talk, I will address how my interest in speech perception has influenced the speech synthesis evaluation research I have been involved in over the last five years. Aspects of evaluation that I will cover include: accent rating, psycholinguistic experiments, artificial personality, comprehension of synthetic speech and the spoofing and voice conversion challenges. I will illustrate the importance of experimental design, as well as discuss some of the open questions in evaluation and remaining unresolved challenges.

Clinical Challenges for Speech Technology
Phil Green, University of Sheffield,Tuesday June 21st, 14:00

Speech Technology is becoming part of everyday life but, paradoxically, it is least effective for those who need it most: people who have problems with speech communication. In this critical review I will cover technology which aims to:

- provide aids for speech professionals (therapists, teacher, pathologists),
- recognise disordered speech and provide voice control,
- help people whose voice is deteriorating,
- restore the power of speech to people whose voice has been lost completely.

I will conclude by outlining the CloudCAST network, whose mission is to provide speech professionals with the tools to build bespoke speech technology using remote resources 'in the cloud'.

Technical programme

Poster Session 1
Monday June 20th, 14:45, Diamond - Workroom 1

- D. A. Baude, B. Potard, M. P. Aylett, G. Pullin, S. Hennig, M. A. Ferreira: "Don't Say Yes, Say Yes: Interacting with Synthetic Speech Using Tonetable"
- P. Bell, S. Renals: "A system for automatic alignment of broadcast media captions using weighted finite–state transducers"
- B. R. Cowan, D. Gannon, J. Walsh, J. Kineen, E. O'Keefe, L. Xie: "Towards Understanding How Speech Output Affects Navigation System Credibility"
- A. Cullen, N. Harte: "Thin Slicing for Speaker Trait Recognition"
- C. Lai, M. Farrus, J. D. Moore: "Automatic Paragraph Segmentation with Lexical and Prosodic Features"
- A. Malinin, R. C. Van Dalen, Y. Wang, K. M. Knill, M. J. F. Gales: "Off–topic Response Detection for Spontaneous Spoken English Assessment"
- R. Milner, T. Hain: "Segment–oriented evaluation of speaker diarisation performance"
- B. Mirheidari and H. Christensen: "Towards the Automatic Conversation Analysis of People with Dementia"
- R. K. Moore: "Can we bridge the 'habitability gap' that arises when everyday users attempt speech–based interaction with 'intelligent' machines?"
- D. Nesbitt, D. Crookes, J. Ming: "Speech–Enhancement for Individuals with Hearing Loss"
- M. Nicolao, H Christensen, S. Cunningham, P. Green, T. Hain: "A Framework for Collecting Realistic Recordings of Dysarthric Speech: the *homeService* Corpus"
- Z. Ommani, M. Hawley, H. Christensen: "Spoken Dialogue System Assisting Older People"
- C. O'Reilly, N. M. Marples, D. J. Kelly, N. Harte: "YIN–bird: Improved Pitch Tracking for Bird Vocalisations"
- B. Potard, M. P. Aylett, D. A. Baude: "Idlak Tangle: An Open Source DNN–Based Parametric Speech Synthesiser"
- S. Ronanki, G. E. Henter, Z. Wu, S. King: "A template–based approach for intonation generation using LSTMs"
- S. Taylor, A. Kato, I. Matthews, B. Milner: "Audio–to–Visual Speech Conversion using Deep Neural Networks"
- C. Valentini-Botinhao, X. Wang, S. Takaki, J. Yamagishi: "Speech Enhancement for a Noise–Robust Text–to–Speech Synthesis System using Deep Recurrent Neural Networks"
- J. A. Vasilakes, H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill: "Speech Recognition and Keyword Spotting Performance Analysis Across Languages"
- P. Weber, L. Bai, S. Houghton, P. Jancovic, M. Russell: "Progress on Phoneme Recognition with a Continuous–State HMM"
- H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, P. Szczpaniak: "Fast, Compact and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices".
- A. Zermini, Y. Yu, Y. Xu, M. D. Plumpley, W. Wang: "Sparse Deep Neural Networks for Audio Source Separation"

Don't Say Yes, Say Yes: Interacting with Synthetic Speech Using Tonetable

David A. Baude[1], *Blaise Potard*[1], *Matthew P. Aylett*[1,2], *Graham Pullin*[3], *Shannon Hennig*[4], *Marilia A. Ferreira*[3]

[1]CereProc Ltd., United Kingdom
[2]The Centre for Speech Research, University of Edinburgh, United Kingdom
[3]Duncan of Jordanstone College of Art & Design, University of Dundee, United Kingdom
[4]Inclusive Communication LTD, New Zealand

{dave,blaise,matthewa}@cereproc.com
g.pullin@dundee.ac.uk, shannon@inclusive-communication.co.nz, mariliaferreiraa@gmail.com

Abstract

This demo is not about what you say but how you say it. Using a tangible system, Tonetable, we explore the shades of meaning carried by the same word said in many different ways. The same word or phrase is synthesised using the Intel Edison with different expressive techniques. Tonetable allows participants to play these different tokens and select the manner they should be synthesised for different contexts using RFID tag enabled cards. By allowing the users to create their own annotations for the cards, the system provokes participants to think deeply about the meaning behind *yes, oh really*, or *I see*. Designed with the very serious objective of supporting expressive personalisation of AAC devices, but with the ability to produce a playful and amusing experience, Tonetable will change the way you think about speech synthesis and what *yes* really means.

Figure 1: *The tonetable in use*

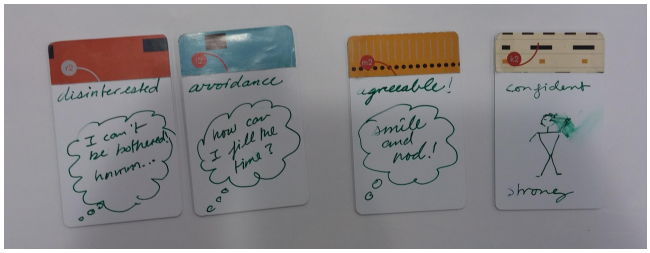


Figure 2: *Sample of users annotations on cards*

A system for automatic alignment of broadcast media captions using weighted finite-state transducers

Peter Bell, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

{peter.bell, s.renals}@ed.ac.uk

1. Abstract

We describe our system for alignment of broadcast media captions in the 2015 MGB Challenge. A precise time alignment of previously-generated subtitles to media data is important in the process of caption generation by broadcasters. However, this task is challenging due to the highly diverse, often noisy content of the audio, and because the subtitles are frequently not a verbatim representation of the actual words spoken. Our system employs a two-pass approach with appropriately constrained weighted finite state transducers (WFSTs) to enable good alignment even when the audio quality would be challenging for conventional ASR. The system achieves an f-score of 0.8965 on the MGB Challenge development set.

2. Summary of method

We apply a two-pass approach with *factor transducers* [1], accepting any sub-string of a supplied string of text. In the first pass, a single grammar transducer, G, is generated for each show. In the second pass, WFSTs are generated dynamically per utterance by selecting surrounding text, and word skips are allowed, similar to [2], giving robustness to deletions (words in the text but not spoken).

```
he loves your ******** ** PICTURE he thinks ****** YOU'LL do ********** well in milan

he loves your PICTURES SO MUCH    he thinks YOU'RE GONNA  do INCREDIBLY well in milan
```

Figure 1: *Example human-generated TV captions (above) compared with verbatim transcription (below)*

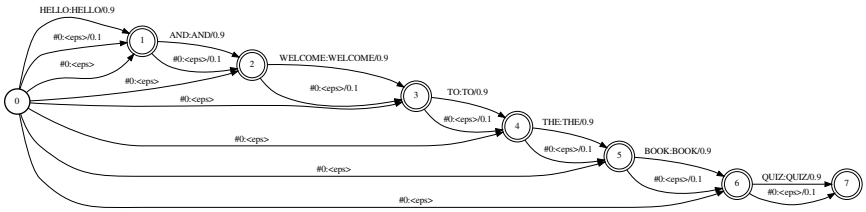


Figure 2: *Factor transducer with optional skips, prior to determinisation. The #0 symbols map to a short-pause (tee) model*

3. References

[1] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. ICASSP*, 2009.

[2] T. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. Interspeech*, 2006.

Towards Understanding How Speech Output Affects Navigation System Credibility

Benjamin R. Cowan, Derek Gannon, Jenny Walsh, Justin Kinneen, Eanna O'Keefe, Linxin Xie

University College Dublin, Ireland
benjamin.cowan@ucd.ie

1. Abstract

Navigation systems are widely used yet little is understood about how aspects of the interaction impact the user experience. Our work focuses on the speech output of these systems, exploring how accent and destination errors delivered in the speech affect user credibility judgements. A small scale study was run whereby Irish participants followed directions given by a computer using either an Irish (CereVoice Caitlin) or American (CereVoice Hannah) synthesised voice to eight destinations on an interactive map of an Irish town. The directions given by the computer varied depending on whether they guided people to the correct destination all of the time (accurate condition) or only half of the time (inaccurate condition). Findings showed that destination errors significantly affected user trust [F (1,34)=11.36, p=.002] and competence assessment [F (1,34)=19.70, p<.001]. The participants also rated the system using Irish accented speech as more trustworthy than the system using American accented speech [F (1,34)=5.35, p=.027]. These findings seem to be an example of the *similarity attraction effect*, whereby we judge communicative systems that are more similar to ourselves more positively [1]. Although previous work has ruled out the role of knowledge assumptions due to accent explaining such findings [2] our planned future work is looking to test this claim in a navigation context by varying the geographical locations being used in a similar experimental set up.

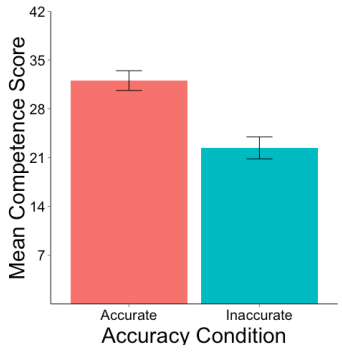


Figure 1- Mean competence score (with standard error) for each accuracy condition

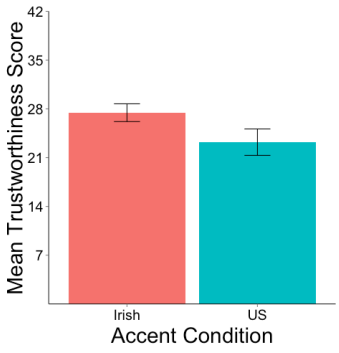


Figure 2- Mean trustworthiness score (with standard error) for each accent condition

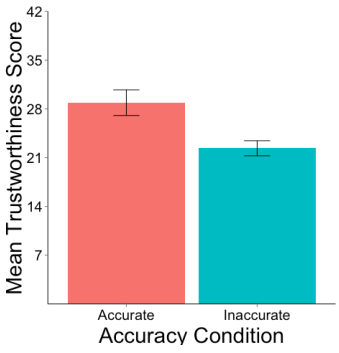


Figure 3- Mean trustworthiness score (with standard error) for each accuracy condition

2. References

[1] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*, New edition. Cambridge University Press, 1998.

[2] N. Dahlbäck, Q. Wang, C. Nass, and J. Alwin, 'Similarity is More Important Than Expertise: Accent Effects in Speech Interfaces', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2007, pp. 1553–1556.

.

Thin Slicing for Speaker Trait Recognition

Ailbhe Cullen, Naomi Harte

Dept. of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

cullena3@tcd.ie, nharte@tcd.ie

1. Abstract

The wealth of data and meta-data which has built up around the TED Talks database has made it a valuable resource for the speech processing and machine learning communities. In particular, the availability of affective user labels (for example inspiring, obnoxious, persuasive) enable large-scale analysis of paralinguistic traits. In this talk we will explore the classification of three user-generated labels, funny, inspiring, and persuasive, for TED talks. Since the talks are collected from existing videos, we have no control over recording set-up, video-editing effects (shot cuts, zooms, slides, etc.), video quality, or audience noise. This presents challenges to automatic audio and video processing routines, but also provides us with new sources of information. In this chapter we will exploit audience feedback, in the form of laughter and applause, to improve classification performance on the three labels. The variable length of the recordings from 3 - 12 minutes, also poses a new challenge. The majority of existing paralinguistic studies perform both annotation and classification at the same temporal level. We explore a thin slicing technique, popular in psychological research but rarely used in computational work, to map long talks to single word labels. We demonstrate that provided slice position is well chosen, we can achieve high recognition accuracies using very short clips from each talk.

Automatic Paragraph Segmentation with Lexical and Prosodic Features

Catherine Lai[1]*, Mireia Farrús*[2]*, Johanna D. Moore*[1]

[1]School of Informatics, University of Edinburgh, Edinburgh, UK
[2]TALN Research Group, DTIC, Universitat Pompeu Fabra, Barcelona, Spain
clai@inf.ed.ac.uk, mireia.farrus@upf.edu, j.moore@ed.ac.uk

1. Introduction

As long-form spoken documents become more ubiquitous in everyday life, so does the need for automatic discourse segmentation in spoken language processing tasks. Although previous work has focused on broad topic segmentation, detection of finer-grained discourse units, such as paragraphs, is highly desirable for presenting and analyzing spoken content. To better understand how different aspects of speech cue these subtle discourse transitions, we investigate automatic paragraph segmentation of TED talks. In particular, we examine how lexical and prosodic features that help high level topic segmentation work at the paragraph level.

2. Experiments

We built paragraph boundary detectors based on a set of 1365 TED (Technology, Entertainment, Design) talks published before 2014. Talks are 15 minutes long on average and vary greatly in content and style (1156 unique speakers). We extract surface, syntactic, and language model features based on previous work on for paragraph segmentation of text. We also measure coherence based using features from unsupervised topic modelling and lexical chain scores. In terms of prosody, we extract F0, intensity and timing/pause features over sentence units. We build lexical and prosodic paragraph segmenters using Support Vector Machines, AdaBoost, and Bi-directional Long Short Term Memory (BLSTM) recurrent neural networks. The 10-fold cross-validation results are evaluated using standard segmentation metrics: P_k, WindowDiff, and k-κ.

3. Results and Conclusions

In general, results indicated that induced cue words and supra-sentential prosodic features outperform features based on topical coherence, syntactic form and complexity. However, our best performance is achieved by combining a wide range of individually weak lexical and prosodic features, with the sequence modelling BLSTM generally outperforming the other classifiers by a large margin. Futher BLSTM experiments indicate that performing lexical and prosodic fusion at intermediate (i.e. hidden layer) levels produced better results that decision level or feature level fusion. That is, we find that models that allow lower level interactions between different feature types produce better results than treating lexical and prosodic contributions as separate, independent information sources. In the future, we plan to investigate the relationship between cue words, prosody, and hierarchical structure for automatically segmentation, and to better shed light on the relationship between topic and rhetorically based notions of discourse structure.

Feature set	SVM	AdaBoost	BLSTM
lm	0.00	0.09	0.04
syntax	0.02	0.02	0.11
pos	0.02	0.02	0.13
bow	0.08	0.09	0.17
cw	0.09	0.07	0.17
surface	*0.11*	*0.14*	*0.24*
dur	0.10	0.13	0.13
prosody	*0.11*	*0.19*	*0.21*
lex.coh	0.07	0.10	0.10
lex.base	0.13	0.16	0.25
lex.all	*0.14*	*0.17*	*0.28*
cw+bow	0.10	0.09	0.21
cw+prosody	0.13	0.21	0.28
lex.all+prosody	**0.17**	**0.26**	**0.31**

Table 1: *k-κ (Niekrasz and Moore 2010) results for SVM, AdaBoost, BLSTM models (larger values are better).*

Off-topic Response Detection for Spontaneous Spoken English Assessment

Andrey Malinin, Rogier C. Van Dalen, Yu Wang, Kate M. Knill, Mark J. F. Gales
University of Cambridge, Department of Engineering
Trumpington St, Cambridge CB2 1PZ, UK
{am969, yw396, kmk1001, mjfg}@cam.ac.uk

Abstract

Automatic spoken language assessment systems are becoming increasingly important to meet the demand for English second language learning. This is a challenging task due to the high error rates of, even state-of-the-art, non-native speech recognition. Consequently current systems primarily assess fluency and pronunciation. However, content assessment is essential for full automation. As a first stage it is important to judge whether the speaker responds on topic to test questions designed to elicit spontaneous speech. Standard approaches to off-topic response detection assess similarity between the response and question based on bag-of-words representations. While simple, these methods lose important sequence information and scale poorly with the amount of training data, which places practical limits on the performance of such systems.

An alternative framework based on Recurrent Neural Network Language Models (RNNLM) is proposed in this paper. The RNNLM is adapted to the topic of each test question. It learns to associate example responses to questions with points in a topic space constructed using these example responses. Classification is done by ranking the topic-conditional posterior probabilities of a response. The RNNLMs associate a broad range of responses with each topic, incorporate sequence information and scale better with additional training data, unlike standard methods. Thus, inference times are unaffected by training data size.

Experiments conducted on data from the Business Language Testing Service (BULATS). The usage scenario is to classify topic relevance based on ASR transcriptions of speaker responses. Thus, to create a model which matches the test data, the system must be trained on ASR transcriptions. Speech recognition is used to generate transcriptions of a large portion of untranscribed BULATS data. Two data sets are generated - a smaller 490 speaker data set and a 10004 speaker datasets, which is 20x larger. In experiments these automatically derived transcriptions are used to train the both the standard and RNNLM off-topic response detection system. Despite an average word error rate of 30%, these systems perform well. Evaluation the systems on professional manual transcriptions of the test data shows a very marginal increase in performance, showing that the systems are robust to ASR. errors.

Two forms of closed-set experiments are conducted - topic detection and off-topic response detection. In the former, the system have to correctly classify the topic of a response. In the latter experiment, the systems have to identify whether a response is invalid for a particular topic. The test data does not contain real off-topic responses. Thus, valid responses to other questions as selected to be used as invalid responses to a particular question of interest. Two selection strategies were investigated - responses could be chosen either from any BULATs test section, called the `Naive` strategy or only from the same section as the question of interest, called the `Directed` Strategy. Overall performance is assessed by plotting False Rejection rate vs. False Acceptance rate on a Receiver-Operator Curve (ROC). In all cases, the RNNLM architecture outperforms standard approaches when trained on equal amounts of data. When the RNNLM is trained on 20x more data, its performance is significantly improved.

Segment-oriented evaluation of speaker diarisation performance

Rosanna Milner, Thomas Hain

Speech and Hearing Research Group, University of Sheffield, UK

rmmilner2, t.hain@sheffield.ac.uk

1. Abstract

High performance diarisation is a necessity for a variety of applications, and the task has been studied extensively in the context of broadcast news and meeting processing. Upon introduction of the task in NIST led evaluations, diarisation error rate (DER) was introduced as the standard metric for evaluation, and it has been consistently used to compare systems ever since. DER is a frame based metric that does not penalise for producing many short segments. However, practical systems that require diarisation input are typically not able to cope well with such artefacts. For example it was repeatedly shown that DER and ASR word error rate do not correlate well. In this paper we illustrate the need for an alternative metric focussing on segments, instead of duration or boundaries only. We propose a segment based F-measure, which specifically addresses issues such as reference errors, matching start and end boundaries, and speaker pairing. The performance of the metric is analysed in the context of state-of-the-art systems and compared with other existing metrics. It is shown to give a deeper insight into the segmentation quality over the standard metrics, and thus better value for to understand impact on follow on tasks such as ASR.

Towards the Automatic Conversation Analysis of People with Dementia

Bahman Mirheidari and Heidi Christensen
Department of Computer Science, University of Sheffield, Sheffield, UK
{bmirheidari2,heidi.christensen}@sheffield.ac.uk

Abstract

The early diagnosis of dementia is a major concern for governments and health services. The production and reception of language are complex achievements which may be affected early by neurodegenerative disorders (ND). For this reason, tests based on comprehension and word fluency form a key part of routine screening tests for dementia. A recent study using conversation analysis (CA) has demonstrated that language and communication problems may also be picked up during interactions between patients presenting with memory problems and neurologists, and that this can be used to differentiate between patients with (progressive) ND and those with (nonprogressive) functional memory disorders (FMD). The main objective of the current study was to explore whether the differential diagnostic analysis of the interaction between patient and doctor could be automated and whether a computer program could be trained to differentiate verbatim transcripts of interactions reliably into those relating to patients with ND and FMD.

Verbatim transcripts of conversations between neurologists and patients initially presenting with memory problems to a specialist clinic (22 with FMD, and 18 with ND) were produced. All patients had received "gold standard" medical diagnosis on the basis of expert assessment, brain imaging and detailed neuropsychological testing. A range of acoustic, syntactic, semantic (pragmatic) and visual features were automatically extracted from the transcripts and used to train a set of classifiers aiming to distinguish between transcripts of patients with ND or FMD. Using the cross validation leave-one-out technique, the classifiers were trained and the diagnostic accuracy of the process evaluated.

The mean rate of correctly classifying to either ND or FMD was 84%. However, some individual classifiers produced higher scores (e.g. Linear SVM with 95% accuracy). Using only the ten best features rather than all features lead to further improvement with a mean correct classification score of 87%. This pilot study provides proof-of-principle that an automatic conversation analysis tool can approximate formal qualitative CA profiling from outpatient consultations between neurologists and patients aiming to distinguish between ND and FMD. At this pilot-stage, the results are gained by a semi-automatic approach because manually produced (verbatim) transcripts were used for the computerised analysis. A further automated approach would involve automated speech recognition. Our findings suggest that it may be feasible to develop an automated diagnostic screening tool, based on patients' interactional behaviour, capable of identifying patients in the early stages of ND.

Can we bridge the 'habitability gap' that arises when everyday users attempt speech-based interaction with 'intelligent' machines?

Roger K. Moore

Speech and Hearing Research Group,
Dept. Computer Science, University of Sheffield,
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
r.k.moore@sheffield.ac.uk

1. Abstract

The release in 2011 of *Siri*, Apple's voice-based personal assistant for the iPhone, signalled a step change in the public perception of spoken language technology. For the first time, a significant number of everyday users were exposed to the possibility of using their voice to enter information, navigate applications or pose questions - all by speaking to their mobile device. Of course, voice dictation software had been publicly available since the release of *Dragon Naturally Speaking* in 1997, but such technology only found success in niche market areas for document creation (by users who could not or would not type). In contrast, *Siri* appeared to offer a more general-purpose interface that thrust the potential benefits of automated speech-based interaction into the forefront of the public's imagination. By combining automatic speech recognition and speech synthesis with natural language processing and dialogue management, *Siri* promoted the possibility of a more *conversational* interaction between users and smart devices [1]. As a result, competitors such as *Google Now* and Microsoft's *Cortana* soon followed.

Of course, it is well established that, while voice-based personal assistants such as *Siri* are now very familiar to the majority of mobile device users, their practical value is still in doubt. This is evidenced by the preponderance of videos on *YouTube*™ that depict humorous rather than practical uses; it seems that people give such systems a try, play around with them for a short while and then go back to their more familiar ways of doing things. Indeed, this has been confirmed by a recent survey of users from around the world which showed that only 13% of the respondents used a facility such as *Siri* every day, whereas 46% had tried it once and then given up (citing inaccuracy and a lack of privacy as key reasons for abandoning it) [2].

This lack of serious take-up of voice-based personal assistants could be seen as the inevitable teething problems of a new(ish) technology, or it could be evidence of something more deep-seated. This paper argues that there is a *habitability gap* [3] caused by the inevitable mismatch between the capabilities and expectations of human users and the features and benefits provided by contemporary technology. Suggestions are made as to how such problems might be mitigated, but a more worrisome question emerges: "*Is spoken language all-or-nothing*"? The answer, based on contemporary views on the special nature of (spoken) language [4], is that there may indeed be a fundamental limit to the interaction that can take place between mismatched interlocutors (such as humans and machines) [5]. However, it is concluded that interactions between native and non-native speakers, or between adults and children, or even between humans and dogs, might provide critical inspiration for the design of future speech-based interaction with 'intelligent' machines.

2. References

[1] Pieraccini, R. (2012). *The Voice in the Machine*. MIT Press, Cambridge, MA.

[2] Moore, R. K., Li, H., & Liao, S.-H. (2016). Progress and prospects for spoken language technology: what ordinary people think. In *INTERSPEECH*. San Francisco, CA.

[3] Philips, M. (2006). Applications of spoken language technology and systems. In M. Gilbert & H. Ney (Eds.), *IEEE/ACL Workshop on Spoken Language Technology (SLT)*.

[4] Scott-Phillips, T. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan.

[5] Moore, R. K. (2016). Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *International Workshop on Spoken Dialogue Systems (IWSDS)*. Saariselk, Finland.

Speech-Enhancement for Individuals with Hearing Loss

David Nesbitt, Danny Crookes, Ji Ming

Queen's University Belfast, Belfast, BT3 9DT

dnesbitt03@qub.ac.uk, d.crookes@qub.ac.uk, j.ming@qub.ac.uk

1. Abstract

Research on Corpus-Based Speech Enhancement algorithms has been pioneered at Queen's University Belfast, seeking to exploit the long-term lingua-acoustic characteristics of speech, in conjunction with the short-term shape of speech segments, to best estimate and enhance speech in noise using a clean-speech corpus. This first took the form of the Longest Matching Segment (LMS) algorithm [1] and has recently evolved into an approach named 'wide matching' [2]. Wide matching finds an optimal estimate for up to a sentence of speech by first matching segments of frames from the corpus and then iteratively reselecting each segment to maximize the intelligibility of the sentence to a speech recognizer. An advantage of this approach is that it makes no assumptions about any noise and, as such, is attractive to real-world speech enhancement, which must deal with non-stationary noise.

While the current prototype version will be computationally expensive and only suitable for offline processing (due to its requirement to have an entire utterance), it is thought that a real-time version could be of use to individuals with hearing loss, with a longer-term goal of adapting the technology to hearing aids. We therefore plan to develop a version of the algorithm that functions in real-time on portable hardware, perhaps with algorithmic compromises. To date we have achieved promising results on a basic segment matching algorithm with output of reasonable quality at a speed of 2.5x faster than real-time on Raspberry Pi 2 test hardware, exploiting its GPU. Work will soon proceed to how best to model the long-term characteristics of speech in a computationally efficient manner.

To achieve these speeds the size of the corpus search space needs to be reduced. Wide matching currently searches sequentially through a clean-speech corpus (Figure 1, left). This search space can be significantly reduced by clustering the corpus into a search tree. Using the search tree, the algorithm can find the best matching segment at the root level, search the child segments of that best matching segment, and so on until a final childless segment is found (Figure 1, right). This enables the system to use much larger corpora than would otherwise be possible in the time. This approach has opened up the whole area of corpus engineering, which will be a key component of our research.

Once the real-time version of the segment matching system is stable for general purpose CPUs, work can then proceed on accelerating the algorithm on hardware, to leave more time for integrating sentence-wide matching. There are two hardware platforms that we intend to target: FPGAs and GPUs. In particular, implementation of key parts of the algorithm on an FPGA would allow for a low-power portable speech enhancement device, with long-term potential to be adapted into a hearing aid system.

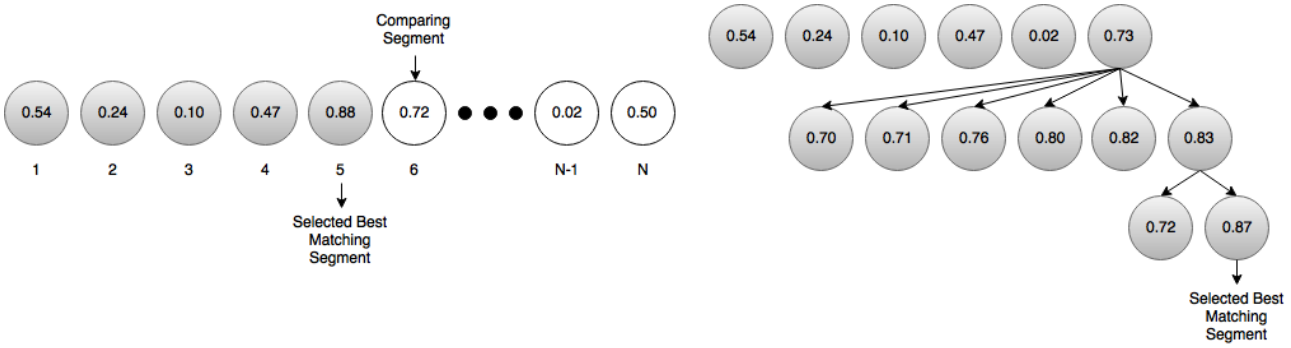


Figure 1: *Left: Sequential search. Right: Cluster-based search.*

2. References

[1] J. Ming and D. Crookes, "An iterative longest matching segment approach to speech enhancement with additive noise and channel distortion," *Computer Speech & Language*, vol. 28, pp. 1269–1286, 2014.

[2] ——, "Wide matching - an approach to improving noise robustness for speech enhancement," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*. Institute of Electrical and Electronics Engineers (IEEE), 2016.

A Framework for Collecting Realistic Recordings of Dysarthric Speech: the *homeService* Corpus

Mauro Nicolao[1], *Heidi Christensen*[1], *Stuart Cunningham*[2], *Phil Green*[1], *and Thomas Hain*[1]

[1]Computer Science, University of Sheffield, UK
[2]Human Communication Sciences, University of Sheffield, UK
{m.nicolao, h.christensen, s.cunningham, p.green, t.hain}@sheffield.ac.uk

1. Abstract

This paper introduces the first release of a new British English speech database, named *homeService corpus*, which has been gathered as part of the *homeService* project. This project aims to help users with speech and motor disabilities to operate their home appliances using voice commands. The audio recorded during such interactions consists of realistic data of speakers with severe dysarthria. The collection of the corpus is motivated by the shortage of realistic dysarthric speech corpora available to the scientific community. The majority of the homeService corpus is recorded in real home environments where voice control is often the normal means by which users interact with their devices.

Seven participants have been recruited, 4 male and 3 female native British speakers with different conditions and age. A total of more than 9 hours of audio has been recorded and annotated from 5 of these participants see Table 1. The corpus consists of two types of speech data:

- **enrolment data, ER**: the user reads lists of the words that they have chosen as commands in their system. To match the acoustic conditions in user's home, the recording takes place with the same hardware and in the same environment in which the system is supposed to function.
- **interaction data, ID**: the user operates the electronic devices in the house with the homeService speech enabled interface. Recording starts after the user presses a switch and the microphone is open for a predefined number of seconds. In contrast with the ER data, the identity and the style of each word produced is not inherently known.

Two subsets have been defined from the ID set. The largest (*ID01train*) is meant to be used along with the entire ER01 data (*ER01train*) for training purposes and the smallest (*ID01test*) represents the set to test ASR system performance.

Table 1: Amount of data collected for homeService v1.0.

Speaker	Type of data	Purpose	Words	Interactions	Duration
F01	ER01train	training	32	97	2'19"
F02	ER01train	training	31	314	11'58"
	ID01train	training	30	314	25'52"
	ID01test	testset	16	85	5'40"
M01	ER01train	training	31	230	6'34"
M02	ER01train	training	31	130	3'16"
	ID01train	training	47	5807	6h29'40"
	ID01test	testset	40	1571	1h44'44"
M03	ER01train	training	12	114	2'47"
	ID01train	training	18	169	11'26"
	ID01test	testset	11	36	3'00"
TOTAL			**131**	**8867**	**9h27'20"**

Speech material and annotations are released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License with the *homeService corpus v1.0* name. Access to the homeService corpus is provided through the dedicated web page at `http://mini.dcs.shef.ac.uk/resources/homeservice-corpus/`. This will also have the most updated description of the data as the collection process is still ongoing.

Along with the details on how the data is organised and how it can be accessed, a brief description of the framework used to make the recordings is provided.

Finally, the performance of the homeService automatic recogniser for dysarthric speech trained with single-speaker (M02) data from the corpus is provided as an initial baseline.

Spoken Dialogue System Assisting Older people

Zahra Ommani, Mark Hawley and Heidi Christensen
School of Health and Related Research University of Sheffield, United Kingdom
Department of Computer Science, University of Sheffield, United Kingdom

{zommani1,mark.hawley,heidi.christensen}@sheffield.ac.uk

Abstract

There are significant number of older people in need of assistance for their daily life activities. However, many of the current assistive technologies (ATs) are not convenient to operate due to complex or confusing interfaces. Speech is a natural means of human communication which can be used in human computer interaction (HCI) to translate a user's intentions through the use of technologies like automatic speech recognition (ASR). Integrating speech technology with AT can provide much simpler interfaces for older people and let them to interact more naturally with devices in their homes.

ASR systems are error prone and they cannot perfectly recognise user utterances and therefore spoken dialogue systems (SDS) are often introduced to further facilitate interactions between users and system. This makes for more natural HCI based on conversations. Although SDS has been the target of mainstream research for some time, with a number of reliable systems in use, there has been less effort put into developing precise and flexible SDS particularly for older people, which is the main concern of this study. Most of the current speech applications are designed for middle-aged or young people, and their ASRs are normally trained by speech collected only from adults or children. A literature review (inspired by the systematic review principles) was carried out to get a better understanding of the important applications, devices, types of speech technologies assisting older people, and the challenges encountered by the developers of those systems.

The review has revealed that the number of studies concerned with using SDS for assisting older people has risen in recent years; likewise using avatars and robots as user interfaces is gaining more popularity amongst the researchers. Moreover, smartphones are often found to be the most convenient devices for developing applications in the caring sector.

We found that designing an appropriate SDS for old users is a complex task with the following major challenges: capturing voice (microphone distance, background noise, quality of recording), dialogue management issues (unexpected responses from the system, ambiguity and complexity of language), TTS quality (synthesised voice causes unpleasant feeling), older people related problems (physical conditions and limitations, chatty people use excessive verbosity while quiet people use short answers, lack of enough previous experience of using computers) and collecting data difficulties (hard to find elderly participants or motivate them for collaboration).

As a further step of this study a SDS will be developed to assist older people in their daily life. Initially the system will allow the elderly to control home devices such as lights and TV and then a calendar will be added to remind them of important events in their daily time schedule. However before designing the real system, the Wizard of OZ (WOZ) approach will be used to collect data from elderly as well as their feedback and suggestions while interacting with the SDS. The information gathered from the WOZ experiments will help us to develop our ultimate SDS.

YIN-bird: Improved Pitch Tracking for Bird Vocalisations

Colm O'Reilly[1], *Nicola M. Marples*[2], *David J. Kelly*[2], *Naomi Harte*[1]

[1] Sigmedia, Department of Electrical & Electronic Engineering, Trinity College Dublin, Ireland
[2]Trinity Centre for Biodiversity Research & Department of Zoology, Trinity College Dublin, Ireland
oreilc16@tcd.ie, nharte@tcd.ie

1. Abstract

Pitch is an important property of birdsong. Accurate and automatic tracking of pitch for large numbers of recordings would be useful for automatic analysis of birdsong. Currently, pitch trackers such as YIN can work with carefully tuned parameters but the characteristics of birdsong mean those optimal parameters can change quickly even within a single song. Work here presents YIN-bird, a modified version of YIN which exploits spectrogram properties to automatically set a minimum fundamental frequency parameter for YIN. This parameter is continuously updated without user intervention. A ground truth dataset of synthetic birdsong with known fundamental frequency is generated for evaluation of YIN-bird. The dataset contains bird whistles, trills and nasal vocalisations. Listener tests from expert birders described the synthetic samples as "sounding like original & can hardly tell it is synthetic". Gross pitch error on whistles and trills were reduced by up to 4%. An analysis of nasal sounds shows the challenge in accurate pitch tracking for this syllable type. A qualitative analysis of other types of syllables, for which a ground truth pitch could not be established, is also discussed.

Figure 1: *UK Speech.*

Idlak Tangle: An Open Source DNN-Based Parametric Speech Synthesiser

Blaise Potard[1], *Matthew P. Aylett*[1,2], *David A. Baude*[1]

[1]CereProc Ltd., United Kingdom
[2]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{blaise,matthewa,dave}@cereproc.com

Abstract

We present a text to speech (TTS) extension to Kaldi - a liberally licensed open source speech recognition system. The system, Idlak Tangle, uses recent deep neural network (DNN) methods for modelling speech, the Idlak XML based text processing system as the front end, and a newly released open source mixed excitation MLSA vocoder included in Idlak. The system has none of the licensing restrictions of current freely available HMM style systems, such as the HTS toolkit. To date no alternative open source DNN systems are available. Tangle combines the Idlak front-end and vocoder, with two DNNs modelling respectively the units duration and acoustic parameters, providing a fully functional end-to-end TTS system.

Experimental results using the freely available SLT speaker from CMU ARCTIC, reveal that the speech output is rated in a MUSHRA test as significantly more natural than the output of HTS-demo, the only other free to download HMM system available with no commercially restricted or proprietary IP. The tools, audio database and recipe required to reproduce the results presented are fully available online at https://github.com/bpotard/idlak.

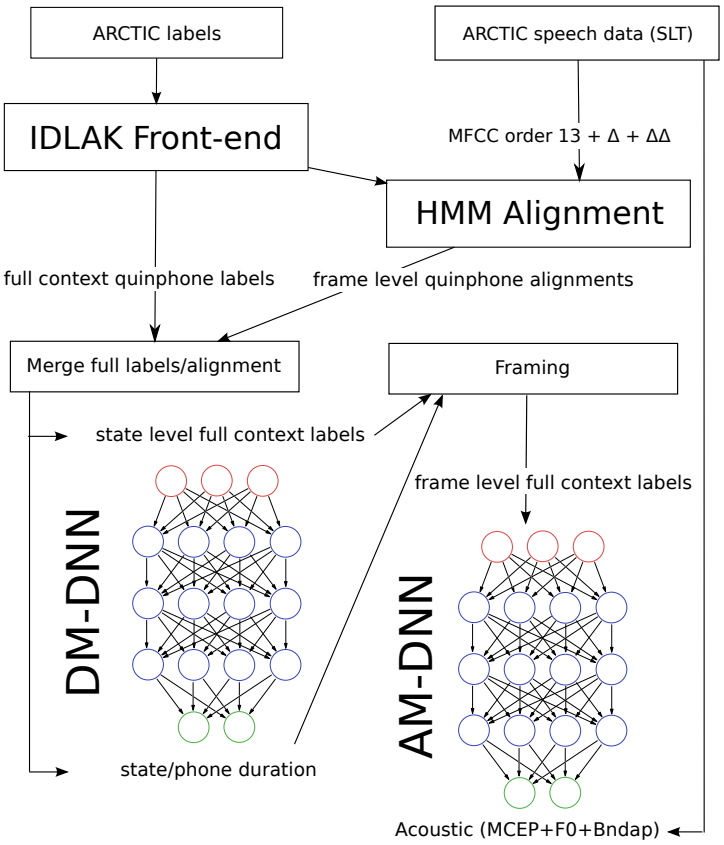


Figure 1: *Tangle DNN training architecture*

A template-based approach for intonation generation using LSTMs

Srikanth Ronanki, Gustav Eje Henter, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK

srikanth.ronanki@ed.ac.uk

1. Abstract

The lack of convincing intonation makes current parametric speech synthesis systems sound dull and lifeless, even when trained on expressive speech data. Typically, these systems predict the fundamental frequency (F0) frame-by-frame using regression models. This approach leads to overly-smooth pitch contours and fails to construct an appropriate prosodic structure across the full utterance. In order to capture and reproduce larger-scale pitch patterns, we propose a classification-based approach to automatic F0 generation, where per-syllable pitch-contour templates (from a small, automatically-learned set) are predicted by a recurrent neural network (RNN). The use of templates mitigates the over-smoothing problem: with only six templates, we can reconstruct pitch patterns observed in the data well (small RMSE). The long memory of RNNs in principle enables the prediction of pitch-contour structure spanning the entire utterance. To construct a complete text-to-speech system, this novel F0 prediction system is used alongside separate LSTMs for predicting phone durations and remaining acoustic features. The objective results are encouraging, but listening tests with oracle reconstructions suggest that further work (beyond a simple smoothing) is necessary to reduce subjective artefacts in the template-based F0 reconstructions.

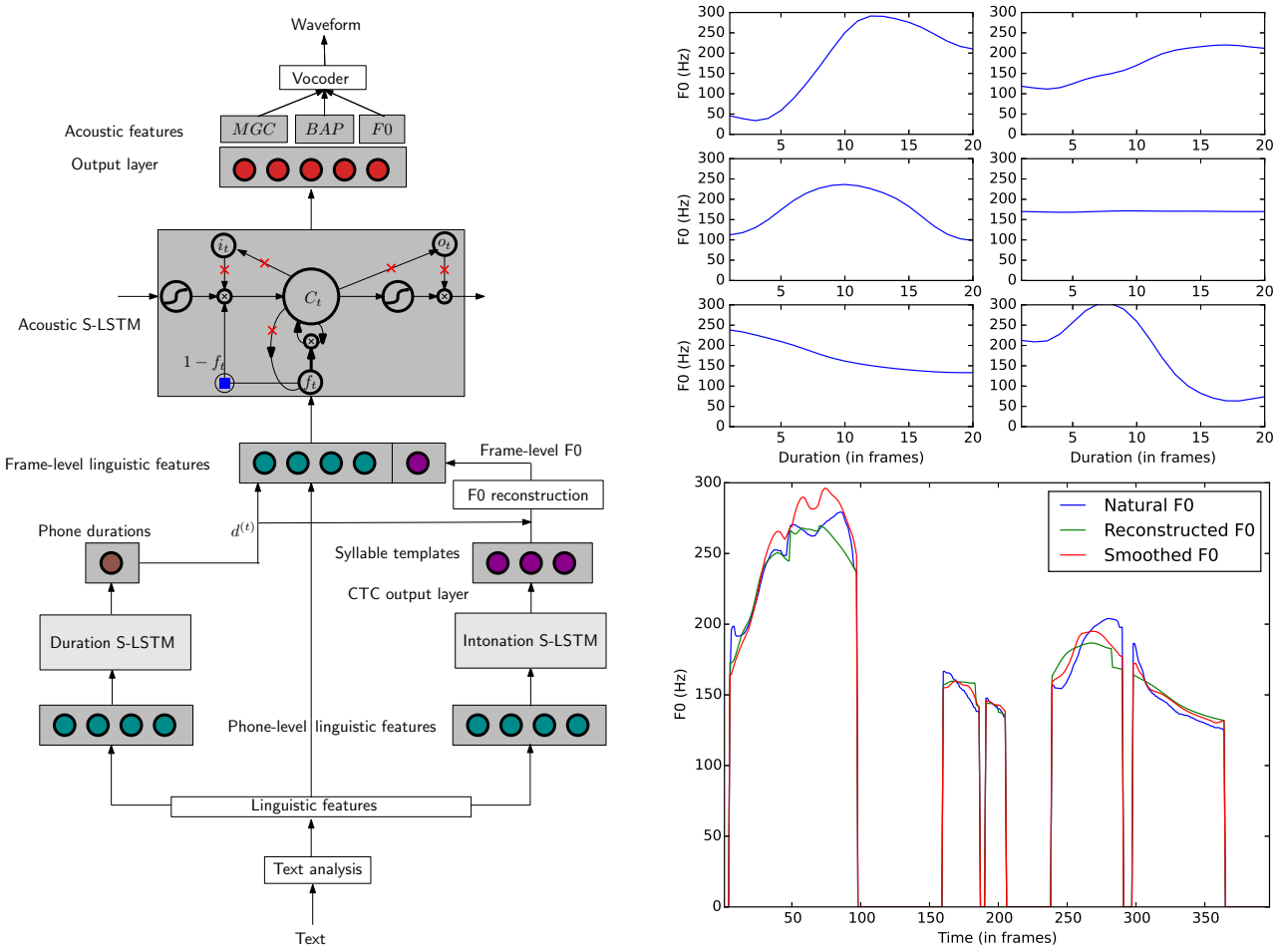


Figure 1: Schematic diagram (left) of the proposed speech synthesis system using a set of six syllable F0 templates (top right). For clarity, only a single LSTM unit is shown. The connections crossed out in red are omitted in the simplified LSTM units [1] used in this work. The bottom right plot shows raw and smoothed F0 contours reconstructed from an oracle template decomposition of natural F0.

2. References

[1] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. ICASSP*, 2016, pp. 5140–5144.

Audio-to-Visual Speech Conversion using Deep Neural Networks

Sarah Taylor[1]*, Akihiro Kato*[1]*, Iain Matthews*[2] *and Ben Milner*[1]

[1]University of East Anglia, Norwich, UK
[2]Disney Research, Pittsburgh, USA

s.l.taylor@uea.ac.uk, akihiro.kato@uea.ac.uk, iainm@disneyresearch.com, b.milner@uea.ac.uk

1. Abstract

We study the problem of audio-to-visual speech conversion (AVSC) which is the task of predicting speech-related facial motion from the acoustic signal with application to fast content creation for animated productions and low-bandwidth multimodal communication.

A variety of approaches have been used to estimate visual features automatically from acoustic speech [1, 2, 3]. A popular approach is to use a form of hidden Markov model (HMM) [4, 5, 6], which use maximum likelihood optimization to predict visual features for given audio. More recently, deep neural networks (DNNs) have been explored for AVSC [3, 7]. DNNs are attractive because they do not impose any Gaussian constraints upon the distribution of the data.

We present a sliding window DNN that learns a mapping from a window of acoustic features (MFCCs) to a window of visual features (active appearance model (AAM) parameters) from a large audio-visual speech dataset. Overlapping visual predictions are averaged to generate continuous, smoothly varying speech animation. This approach has previously worked well for other spatio-temporal sequence prediction tasks [8]. The windowed input *and* output considers carry-over and anticipatory coarticulation in both the acoustic *and* visual modalities and the sliding window predictor alleviates the need for arbitrary smoothing, which is necessary for those methods that predict a single frame at a time [7]. We outperform a baseline HMM inversion approach in both objective and subjective evaluations and perform a thorough analysis of our results.

Specifically, our contributions and key findings can be summarised as follows:

- We propose a full audio-to-visual speech conversion sliding window framework which runs in real-time and requires no phonetic annotation of the input or smoothing of the output.
- We extend conventional DNNs with windowed audio input *and* visual output to account for both acoustic and visual coarticulation effects.
- We investigate the number of MFCCs for AVSC, achieving lowest prediction error when retaining 25 coefficients.
- We investigate the input and output window size on AVSC accuracy and determine that using a 340ms acoustic window to train a three-layer neural network to predict 100ms visual output provides optimal results.
- We show that our method outperforms a baseline HMM inversion approach both objectively and subjectively.
- We explore the effectiveness of acoustic speech for predicting visual speech and discover that facial motion of sibilant fricative and affricate consonants can be predicted most accurately from audio speech, and velar consonants have highest prediction error.

2. References

[1] M. S. Craig, P. van Lieshout, and W. Wong, "A linear model of acoustic-to-facial mapping: Model parameters, data set size, and generalization across speakers," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3183–3190, 2008.

[2] R. Gutierrez-Osuna, P. K. Kakumanu, A. Esposito, O. N. Garcia, A. Bojórquez, J. L. Castillo, and I. Rudomín, "Speech-driven facial animation with realistic dynamics," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 33–42, 2005.

[3] A. Savran, L. M. Arslan, and L. Akarun, "Speaker-independent 3d face synthesis driven by speech and text," *Signal processing*, vol. 86, no. 10, pp. 2932–2951, 2006.

[4] L. D. Terissi and J. C. Gómez, "Audio-to-visual conversion via hmm inversion for speech-driven facial animation," in *Advances in Artificial Intelligence-SBIA 2008*. Springer, 2008, pp. 33–42.

[5] L. Xie and Z.-Q. Liu, "A coupled hmm approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.

[6] X. Zhang, L. Wang, G. Li, F. Seide, and F. K. Soong, "A new language independent, photo-realistic talking head driven by voice only." in *INTERSPEECH*, 2013, pp. 2743–2747.

[7] P. Hong, Z. Wen, and T. S. Huang, "Real-time speech-driven face animation with expressions using neural networks," *Neural Networks, IEEE Transactions on*, vol. 13, no. 4, pp. 916–927, 2002.

[8] T. Kim, Y. Yue, S. Taylor, and I. Matthews, "A decision tree framework for spatiotemporal sequence prediction," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15, 2015, pp. 577–586.

Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks

Cassia Valentini-Botinhao[1], *Xin Wang*[2,3], *Shinji Takaki* [2], *Junichi Yamagishi*[1,2,3]

[1] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] National Institute of Informatics, Japan
[3] SOKENDAI University, Japan

cvbotinh@inf.ed.ac.uk, {wangxin,takaki,jyamagis}@nii.ac.jp

1. Abstract

Quality of text-to-speech voices built from noisy recordings is diminished. In order to improve it we propose the use of a recurrent neural network to enhance acoustic parameters prior to training. We trained a deep recurrent neural network using a parallel database of noisy and clean acoustics parameters as input and output of the network. The database consisted of multiple speakers and diverse noise conditions. We investigated using text-derived features as an additional input of the network. We processed a noisy database of two other speakers using this network and used its output to train an HMM acoustic text-to-synthesis model for each voice. Listening experiment results showed that the voice built with enhanced parameters was ranked significantly higher than the ones trained with noisy speech and speech that has been enhanced using a conventional enhancement system. The text-derived features improved results only for the female voice, where it was ranked as highly as a voice trained with clean speech.

2. Proposed speech enhancement for TTS

Vocoder parameters that describe the source and the filter are extracted from a time frame of the noisy waveform in the same manner as usually done for TTS training. These parameters are then feed to a neural network together with text-derived features that describe the linguistic context of that particular acoustic frame. The network outputs an enhanced set of acoustic parameters that is then used in conjunction with text-derived features to train a text-to-speech acoustic model. Integrating the speech enhancement as a pre-processing stage while directly enhancing the parameters that are going to be used for training the TTS model avoids unnecessary distortions caused by reconstruction of the waveform. The structure could also be seen as a pre-filter (as opposed to a postfilter that acts at the vocoder level at generation time) and in that sense could potentially be used to minimise synthesis errors as well as enhancement errors.

3. Results and discussion

As expected natural speech ranked highest, followed by synthetic speech trained with clean data (CLEAN), while the synthetic voice trained with noisy data (NOISY) ranked worst. Both RNN-based methods performed better than the conventional speech enhancement method (OMLSA). Objective distortion measures calculated over the vocoded parameters showed that the network trained with acoustic and text features (RNN-AT) produces slightly more errors than the one trained with acoustic features only (RNN-A). Listening test scores for the female voice showed a different trend: the synthetic voice trained with speech enhanced using acoustics and text was scored slightly higher and it was not rated significantly different from the voice trained using clean speech. Text-derived features did not improve quality of the synthetic male voice, but for that particular voice all enhancement methods performed worse.

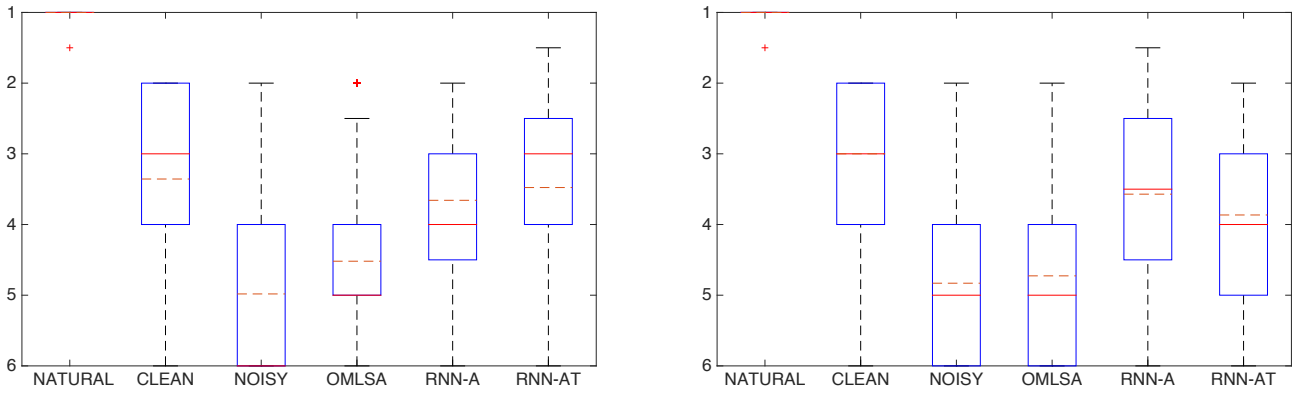


Figure 1: *Rank results of listening experiment with the synthetic female (left) and male (right) voice.*

Speech Recognition and Keyword Spotting Performance Analysis Across Languages

J. A. Vasilakes, H. Wang, A. Ragni, M. J. F. Gales, and K. M Knill

University of Cambridge Department of Engineering

Trumpington Street, Cambridge, CB2 1PZ, UK

{jav39,hw443,ar527,mjfg,kate.knill}@eng.cam.ac.uk

Abstract

There has been recent interest in applying speech processing beyond traditionally used languages. Performance on these languages varies dramatically even using similarly configured systems, trained on similar quantities of data, and run on the same task. A better understanding of the causes of these variations is the first step towards improving and predicting performance for new languages. We present an investigation of these variations through an analysis of the correlation between standard attributes (e.g. size of the phone set, audio quality, language model perplexity) and final performance. Two speech processing tasks are used: automatic speech recognition (ASR) and keyword spotting (KWS). State-of-the-art systems were built for 17 languages from the IARPA Babel programme, selected to cover a range of language families. Correlations between each attribute and final ASR and KWS performance were measured. For most attributes, correlations with observed performance were not significant. However, it was found that a grapheme error rate criterion of ASR systems exhibits a significant correlation to final ASR and KWS performance. This criterion can be rapidly estimated at the early stages of ASR system development. Using this criterion performance predictions were made for 6 held out languages and found to well approximate observed values.

Progress on Phoneme Recognition with a Continuous-State HMM

Philip Weber, Linxue Bai, Steve Houghton, Peter Jančovič and Martin Russell

University of Birmingham, United Kingdom
dr.philip.weber@ieee.org,{lxb190,s.houghton,p.jancovic,m.j.russell}@bham.ac.uk

1. Introduction

Recent progress in automatic speech recognition has predominantly used statistical methods such as Deep Neural Networks (DNNs). Many parameters trained using very large corpora enable accurate, discriminative modeling of distributions over speech features.

Such data-driven training often ignores or contradicts the nature of human speech production and perception: speech generated by the relatively slow, constrained and smooth movement of a small number of articulators in the vocal tract, and features therefore strongly correlated in time and typically exhibiting smooth, slowly-varying dynamics. The cost can be inflexibility when applied to speech from outside the target domain – demonstrated by active research into recognition of many types of 'non-standard' speech such as accented, children, dysarthric (see references in [1]).

It has long been argued [2] that speech features lie on a low-dimensional data manifold embedded in high-dimensional acoustic space, and that modelling this manifold directly would allow recognition to be carried out closer to the original intent, perhaps therefore more robustly to noise and variability. It would also allow the dynamics of the signal to be taken into account. Segmental and dynamical models attempt to model the dynamics of speech more faithfully, but have been hampered by computational complexity.

The Continuous State HMM (CS-HMM) [3] can be cast as a type of segmental model. Its iterative computations avoid some of these problems, and it can be trained on limited data of low dimensionality. Variants have been applied to voiced sounds with formant-type features, and unvoiced sounds using spectral energy features.

We plan to integrate these models into a full recogniser which would probabilistically combine hypotheses from multiple models and heterogeneous views on the data. As an intermediate step, in this work we apply the CS-HMM to an automatically-derived low-dimensional 'bottleneck' (BNF) representation of speech which is valid for all speech sounds [4]. We report promising phoneme recognition results using these bottleneck features (Table 1) and show that the representation is faithful to the assumptions of the CS-HMM (Figure 1).

Features	Model	Sub	Del	Ins	Err	#Parm
39 MFCC+δ+$\delta\delta$	DS-HMM	–	–	–	29.1	1.4e7
9 BNF	DS-HMM	17.8	8.8	2.9	29.4	2.3e5
3 BNF	DS-HMM	24.2	10.8	4.1	39.1	7.6e4
3 Formant	DS-HMM	32.0	18.7	8.6	59.3	7.6e4
3 Formant+δ+$\delta\delta$	DS-HMM	24.3	19.3	5.2	48.9	2.3e5
3 Formant	CS-HMM	35.6	33.4	4.8	73.7	163
3 VTR	CS-HMM	36.2	34.6	3.7	74.6	163
3 BNF	CS-HMM	30.1	14.2	3.6	47.9	163
9 BNF	CS-HMM	22.9	10.2	5.0	38.1	535

Table 1: DS-HMM (HTK [6]) and CS-HMM phone recognition results using formants, VTRs [5] and bottleneck features (BNFs), showing the number of parameters involved.

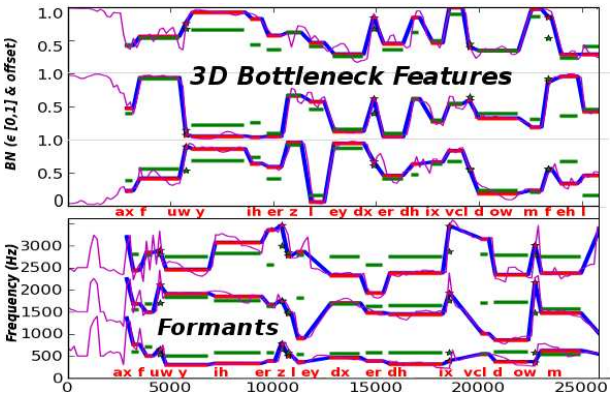


Figure 1: Example CS-HMM recoveries (blue/red), showing realised dwells (red), inventory feature means (green). From top: 3D BNFs (magenta) $\in [0, 1]$, offset to visualise), formants.

2. References

[1] P. Weber, L. Bai, S. Houghton, P. Jančovič and M. J. Russell, "Progress on Phoneme Recognition with a Continuous-State HMM," *In Proc. ICASSP*, 2016, pp. 5850-5854.

[2] G. Fant, *Acoustic Theory of Speech Production*, R. Jakobson and C. H. van Schooneveld, Eds., Mouton, 1970.

[3] C. J. Champion and S. M. Houghton, "Application of Continuous State Hidden Markov Models to a classical problem in speech recognition," *Computer Speech and Language*, 36(1):347–364, 2016.

[4] L. Bai, P. Jančovič, M. Russell, and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," in *Proc. Interspeech 2015*, Dresden, Germany, pp. 583–587.

[5] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. ICASSP 2006*, Toulouse, France, pp. 369–372.

[6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, 2006.

Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices

Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, Przemysław Szczepaniak

Google

{heigazen,agios,nielse,fergus,pszczepaniak}@google.com

1. Abstract

This paper investigated three optimizations of LSTM-RNN-based SPSS for deployment on mobile devices; 1) Quantizing LSTM-RNN weights to 8-bit integers reduced disk footprint by 70%, with no significant difference in naturalness; 2) Using multi-frame inference reduced CPU use by 40%, again with no significant difference in naturalness; 3) For training, using an ϵ-contaminated Gaussian loss function rather than a squared loss function to avoid excessive effects from outliers proved beneficial, allowing for an increased learning rate and improving naturalness. The LSTM-RNN-based SPSS systems with these optimizations surpassed the HMM-based SPSS systems in speed, latency, disk footprint, and naturalness on modern mobile devices. Experimental results also showed that the LSTM-RNN-based SPSS system with the optimizations could match the HMM-driven unit selection TTS systems in naturalness in 13 of 26 languages.

Table 1: *Left: Average latency and total time in milliseconds to synthesize a character, word, sentence, and paragraph by the LSTM-RNN- (`LSTM`) and HMM-based (`HMM`) SPSS systems. Right: Subjective preference scores (%) between the LSTM-RNN- and HMM-based SPSS systems.*

	Latency (ms)		Total (ms)	
Length	`LSTM`	`HMM`	`LSTM`	`HMM`
char.	12.5	19.5	49.8	49.6
word	14.6	25.3	61.2	80.5
sent.	31.4	55.4	257.3	286.2
para.	64.1	117.7	2216.1	2400.8

Language	`LSTM`	`HMM`	No pref.
English (GB)	31.6	28.1	40.3
English (NA)	**30.6**	15.9	53.5
French	**68.6**	8.4	23.0
German	**52.8**	19.3	27.9
Italian	**84.8**	2.9	12.3
Spanish (ES)	**72.6**	10.6	16.8

Table 2: *Subjective preference scores (%) between the LSTM-RNN-based SPSS and HMM-driven unit selection TTS (`Hybrid`) [1] systems. Note that "English (GB)", "English (NA)", "Spanish (NA)", and "Portuguese (BR)" indicate British English, North American English, North American Spanish and Brazilian Portuguese, respectively.*

Language	`LSTM`	`Hybrid`	No pref.
Arabic	13.9	**22.1**	64.0
Cantonese	**25.1**	7.3	67.6
Danish	37.0	**49.1**	13.9
Dutch	29.1	**46.8**	24.1
English (GB)	22.5	**65.1**	12.4
English (NA)	23.3	**61.8**	15.0
French	28.4	**50.3**	21.4
German	20.8	**58.5**	20.8
Greek	**42.5**	21.4	36.1
Hindi	42.5	36.4	21.1
Hungarian	**56.5**	30.3	13.3
Indonesian	18.9	**57.8**	23.4
Italian	28.1	**49.0**	22.9

Language	`LSTM`	`Hybrid`	No pref.
Japanese	**47.4**	28.8	23.9
Korean	**40.6**	25.8	33.5
Mandarin	**48.6**	17.5	33.9
Norwegian	**54.1**	30.8	15.1
Polish	14.6	**75.3**	10.1
Portuguese (BR)	31.4	37.8	30.9
Russian	26.7	**49.1**	24.3
Spanish (ES)	21.0	**47.1**	31.9
Spanish (NA)	22.5	**55.6**	21.9
Swedish	**48.3**	33.6	18.1
Thai	**71.3**	8.8	20.0
Turkish	**61.3**	20.8	18.0
Vietnamese	30.8	30.8	38.5

2. References

[1] X. Gonzalvo, S. Tazari, C.-A. Chan, M. Becker, A. Gutkin, and H. Silen, "Recent advances in Google real-time HMM-driven unit selection synthesizer," in *Proc. Interspeech (submitted)*, 2016.

Sparse Deep Neural Networks for Audio Source Separation

Alfredo Zermini[1], Yang Yu[2], Yong Xu[1], Mark D. Plumbley[1], and Wenwu Wang[1]

1. Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK
Emails: {a.zermini, y.xu, m.plumbley, w.wang}@surrey.ac.uk
2. School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
Email: {nwpuyuy@nwpu.edu.cn}

1. Abstract

Audio source separation aims to extract individual sources from mixtures of multiple sound sources. Many techniques have been developed for this, such as independent component analysis, computational auditory scene analysis, and non-negative matrix factorisation. A method based on Deep Neural Networks (DNNs) and time-frequency masking has been recently developed for binaural audio source separation [1] [2]. In this method, the DNNs are used to predict the Direction Of Arrival (DOA) of the audio sources with respect to the listener. The DOA is then used to generate soft time-frequency masks for the recovery/estimation of the individual audio sources. The DNN consists of two sparse autoencoders and a softmax classifier. The input to the DNN is a vector containing three low-level features, namely: Mixing Vector (MV), Interaural Level Difference (ILD), and Interaural Phase Difference (IPD). These features are all derived from the left and right channels of the binaural recordings.

These method is evaluated on binaural mixtures generated with Binaural Room Impulse Responses (BRIRs) recorded in five different rooms at the University of Surrey [3], representing different level of room reverberations, as used in [4]. The DNNs were trained based on the reverberant sounds generated using these BRIRs. Sound mixtures were obtained by mixing two sound sources convolved with the BRIRs, where one is considered as the target signal, and the other as the interference signal. The target is located at a fixed position, while the interferer is moved around a half-circular grid, in variable positions ranging from −90 degrees to +90 degrees, with steps of 5 degrees. Several DNNs are trained and used to separate the two sound sources where each DNN contains the information from a selected group of frequency bins. Soft-masks derived from the estimated DOAs are used to separate the target from the interferer. The separation results are evaluated in terms of signal to distortion ratio.

2. References

1. Y. Yu, W. Wang, J. Luo, and P. Feng, "Localization based stereo speech separation using deep networks", in *2015 IEEE International Conference on Digital Signal Processing* (DSP), July 2015, pp. 153-157.
2. Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks", *EURASIP Journal on Audio Speech and Music Processing*, 2016:7, 18 pages, DOI 10.1186/s13636-016-0085-x, 2016.
3. C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments", Ph.D. dissertation, University of Surrey, 2011.
4. A. Alinaghi, P. J. B. Jackson, Q. Liu, and W. Wang, Joint mixing vector and binaural model based stereo source separation. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 9, 2014, pp. 1434-1448.

Oral Session
Monday June 20th, 16:20, Diamond - Lecture Theatre 1

- M. A. Figueroa, B. G. Evans: "Coping with loss: Evidence for recovery and lexical effects in the perception of highly lenited Spanish approximants"
- F. Espic, C. Valentini-Botinhao, Z. Wu, S. King: "Waveform generation based on signal reshaping for statistical parametric speech synthesis"
- M. Aylett: "Delighting the User with Speech Synthesis"
- C. Wu, P. Karanasou, M. J. F. Gales, K. C. Sim: "Stimulated Deep Neural Network for Speech Recognition"
- The SUMMA Consortium: "SUMMA – Scalable Understanding of Multilingual Media"

Coping with loss: Evidence for recovery and lexical effects in the perception of highly lenited Spanish approximants

Mauricio A. Figueroa, Bronwen G. Evans

Department of Speech, Hearing and Phonetic Sciences
University College London (UCL), United Kingdom
m.figueroa.12@ucl.ac.uk, bronwen.evans@ucl.ac.uk

1. Abstract

Chilean Spanish spirant approximants [β̞ ð̞ ɣ̞] from /b d g/ display particularly high degrees of lenition in production, which often leads to elision in several phonetic contexts ([1, 2, 3]). Despite lenition being very common, listeners experience no difficulties recovering these units, which raises a series of questions regarding the reliability of the acoustic information cueing for approximants, when and how listeners use complementary cues in perception, how the variability originating from lenition is dealt with during lexical access, and whether this variation is encoded in lexical representations or not. In order to address these issues, synthetic continua from approximant consonant to elision in which both ends were legal Spanish words were built (e.g., from *boga* –['bo.ɣ̞a], "fashionable" or "trendy"– to *boa* –['bo.a], "boa constrictor"), and presented in conditions which varied in the amount of acoustic and semantic cues available, in three perception tasks: phoneme monitoring, identification and discrimination. Results suggested that the expectations from listeners regarding what can be considered normal in natural production and perception had an effect on how each continuum was perceived, with responses closer to categorical perception for those consonants in which the acoustic evidence is more reliable in natural settings (i.e., /g/), and recovery for those in which it is particularly unreliable (i.e., /d/). Adding semantic cues brought all responses closer to categorical distributions, suggesting lexical effects on speech perception. Overall, these results were interpreted as providing evidence for the use of episodic memory in tasks requiring prelexical processing, and top-down feedback from post-lexical levels in tasks in which lexical access is mandatory. These results are thus consistent with both interactive episodic models of lexical access, e.g., Minerva 2 ([4, 5, 6]), and interactive hybrid models, e.g., POLYSP ([7, 8]) and Goldinger's CLS ([9]), in which listeners are thought to store detailed memory representations, and in which top-down feedback is possible.

2. References

[1] Cepeda, G. (1991). *Las consonantes de Valdivia*. Valdivia: Universidad Austral de Chile.
[2] Cepeda, G. (2001). Estudio descriptivo del español de Valdivia, Chile. *Estudios Filológicos*, *36*, 81-97.
[3] Pérez, H. E. (2007). Estudio de la variación estilística de la serie /b-d-g/ en posición intervocálica en el habla de los noticieros de la televisión chilena. *Estudios de Fonética Experimental*, *16*, 228-259.
[4] Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96-101.
[5] Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411-428.
[6] Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251.
[7] Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3), 373-405.
[8] Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics*, *13*, 99-188.
[9] Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th international congress of phonetic sciences* (pp. 49-54).

Waveform generation based on signal reshaping for statistical parametric speech synthesis

Felipe Espic, Cassia Valentini-Botinhao, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
felipe.espic@ed.ac.uk, cvbotinh@inf.ed.ac.uk, zhizheng.wu@ed.ac.uk, Simon.King@ed.ac.uk

1. Abstract

We propose a new paradigm of waveform generation for Statistical Parametric Speech Synthesis (SPSS) that is based on neither source-filter separation nor sinusoidal modelling. We suggest that one of the main problems of current vocoding techniques is that they perform an extreme decomposition of the speech signal into source and filter, which is an underlying cause of ``buzziness", ``musical artifacts", or `` muffled sound" in the synthetic speech. The proposed method avoids making unnecessary assumptions and decompositions as far as possible, and uses only the spectral envelope and F0 as parameters. Pre-recorded speech is used as a base signal, which is ``reshaped" to match the acoustic specification predicted by the statistical model, without any source-filter decomposition. A detailed description of the method is presented, including implementation details and adjustments. A complete diagram of a system including the proposed method is shown in Figure 1. Subjective listening test evaluations of complete DNN-based text-to-speech systems were conducted for two voices: one female and one male. The results show that the proposed method tends to outperform the state-of-the-art standard vocoder STRAIGHT, whilst using fewer acoustic parameters.

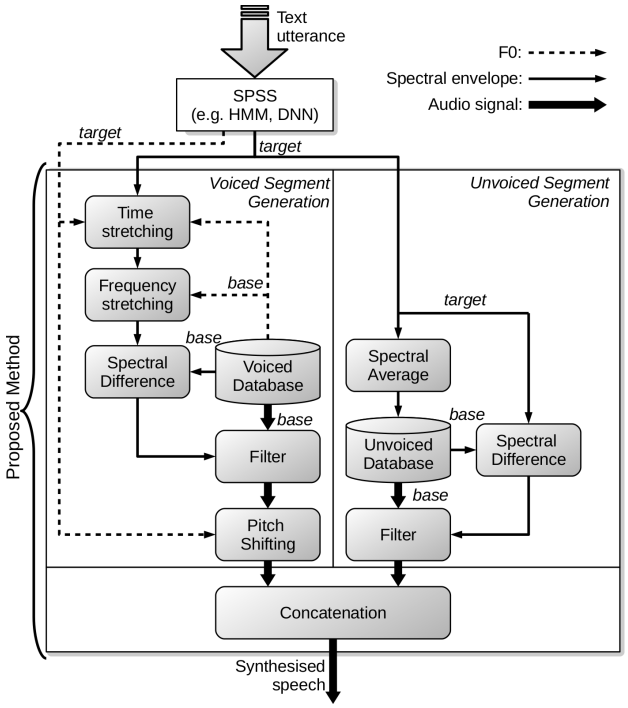


Figure 1: A SPSS system including the proposed method for waveform generation.

Delighting the User with Speech Synthesis

Matthew Aylett

CereProc Ltd., United Kingdom
matthewa@cereproc.com

1. Abstract

We all know there is something special about speech. Our voices are not just a means of communicating, although they are superb at communicating, they also give a deep impression of who we are. They can betray our upbringing, our emotional state, our state of health. They can be used to persuade and convince, to calm and to excite. Speech synthesis technology offers a means to engage the user, to personify an interface, to add delight to human computer interaction. In this talk I will present speech synthesis work that supports social interaction through the use of emotion, personalization and audio design, I will relate this technology to requirements in dialogue systems, eyes-free data aggregation and audio interfaces, and I will discuss the challenges the technology faces for a pervasive, eyes-free future.

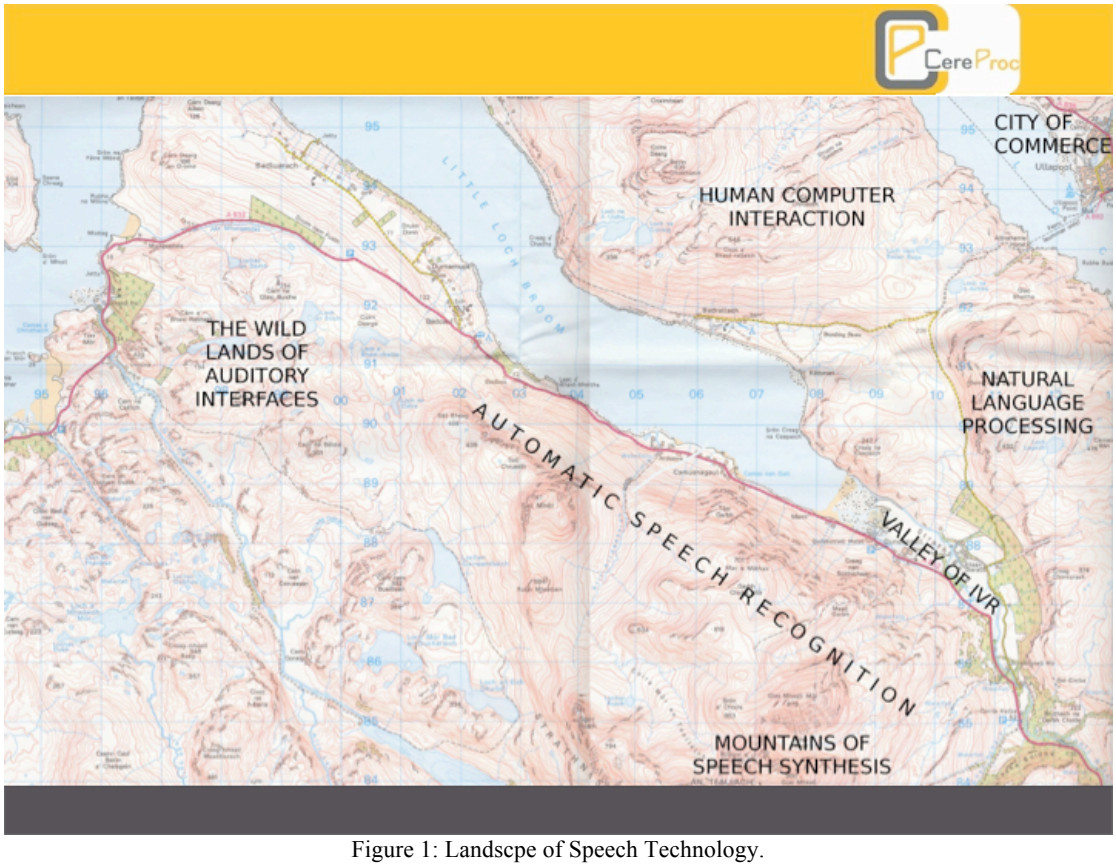


Figure 1: Landscpe of Speech Technology.

Stimulated Deep Neural Network for Speech Recognition

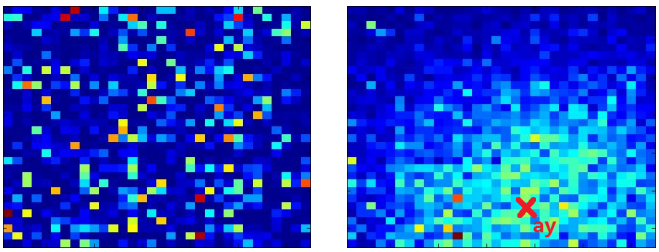
Chunyang Wu[1], *Penny Karanasou*[1], *Mark J.F. Gales*[1], *Khe Chai Sim*[2]

[1]University of Cambridge
[2]National University of Singapore
{cw564, pk407, mjfg}@eng.cam.ac.uk, simkc@comp.nus.edu.sg

1. Abstract

Deep neural networks (DNNs) and deep learning approaches yield state-of-the-art performance in a range of tasks, including speech recognition. However, the parameters of the network are hard to analyze, making network regularization and robust adaptation challenging. Stimulated training [1, 2] has recently been proposed to address this problem by encouraging the node activation outputs in regions of the network to be related: A phone (or grapheme) dependent prior distribution is defined over the normalized activation function outputs for each of the layers, which regularizes activation functions in the same locality to have similar normalized outputs.

The stimulated training aids visualization of the network. As shown in Figure 1, due to the arbitrary ordering issue, there is no stimulated pattern in the grid of the unstimulated DNN. However on the stimulated one, the activation grid nicely corresponded to the stimulating pattern: the nodes around the location of the phone "ay" echoed higher activation values. Meanwhile, it also has the



(a) Unstimulated (b) Stimulated

Figure 1: Comparison of unstimulated and stimulated DNN activations of one hidden layer on an "ay" frame.

potential to improve regularization and adaptation. This research investigates stimulated training of DNNs for both of these options. These schemes take advantage of the smoothness constraints that stimulated training offers. In adaptation, a smoothing technique of the learning hidden unit contributions (LHUC) is put forward. The adapted activation is smoothed by its spatial neighbor activations in order to reinforce the robustness of adaptation.

The approaches are evaluated on several large vocabulary speech recognition tasks: a U.S. English broadcast news (BN) task and conversational telephone speech tasks of seven languages from the IARPA Babel program. Stimulated DNN training acquires consistent performance gains on both tasks over unstimulated baselines. On the BN task, the proposed smoothing approach is also applied to rapid adaptation, again outperforming the standard adaptation scheme. The MPE systems on the BN task are summarized in Table 1. The

System	Dev03	Eval03
MPE	11.6	10.1
+LHUC	11.2	9.8
MPE-Stimu	11.2	9.8
+LHUC	10.9	9.5
+regLHUC	**10.6**	**9.4**

Table 1: MPE Utterance-Level Adaptation on Broadcast News.

speaker-independent (SI) MPE stimulated DNN (MPE-Stimu) outperformed the SI unstimulated MPE baseline. The regularized LHUC (+regLHUC) on the stimulated system achieved the best performance, reducing the WER up to 5% relatively in contrast with the SI MPE stimulated system.

2. References

[1] S. Tan, K. C. Sim, and M. Gales, "Improving the interpretability of deep neural networks with stimulated learning," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 617–623.

[2] C. Wu, P. Karanasou, M. Gales, and K. C. Sim, "Stimulated deep neural network for speech recognition," *Interspeech*, 2016.

SUMMA
Scalable Understanding of Multilingual Media

The SUMMA Consortium

http://summa-project.eu

ABSTRACT

Media monitoring enables the global news media to be viewed in terms of emerging trends, people in the news, and the evolution of story-lines. The massive growth in the number of broadcast and Internet media channels means that current approaches can no longer cope with the scale of the problem.

The aim of SUMMA is to significantly improve media monitoring by creating a platform to automate the analysis of media streams across many languages, to aggregate and distill the content, to automatically create rich knowledge bases, and to provide visualisations to cope with this deluge of data.

SUMMA has six objectives:

1. Development of a scalable and extensible media monitoring platform;
2. Development of high-quality and richer tools for analysts and journalists;
3. Extensible automated knowledge base construction;
4. Multilingual and cross-lingual capabilities;
5. Sustainable, maintainable platform and services;
6. Dissemination and communication of project results to stakeholders and user group.

Achieving these aims will require advancing the state of the art in a number of technologies: multilingual stream processing including speech recognition, machine translation, and story identification; entity and relation extraction; natural language understanding including deep semantic parsing, summarisation, and sentiment detection; and rich visualisations based on multiple views and dealing with many data streams.

The project will focus on three use cases:

1. External media monitoring - intelligent tools to address the dramatically increased scale of the global news monitoring problem;
2. Internal media monitoring - managing content creation in several languages efficiently by ensuring content created in one language is reusable by all other languages;
3. Data journalism.

The outputs of the project will be field-tested at partners BBC and DW, and the platform will be further validated through innovation intensives such as the BBC News-Hack.

The SUMMA project commenced in 2016 for three years, and the project partners comprise: (i) University of Edinburgh (UK); (ii) UCL (UK); (iii) Idiap Reseach Institute (CH); (iv) Priberam (PT); (v) LETA (LV); (vi) BBC (UK); (vii) Deutsche Welle (DE); (viii) University of Sheffield (UK); (ix) QCRI (QT).

Poster Session 2
Tuesday June 21st, 10:00, Diamond - Workroom 1

- N. Alghamdi, S. Maddock, G. J. Brown, J. Barker: "Simulating the Visual Lombard Effect"
- S. Al–Hameed, M. Benaissa, H. Christensen: "Towards simple and robust audio–based detection of biomarkers for Alzheimer's disease"
- A. Ali, S. Khurana, N. Dehak, P. Cardinal, S. H. Yella, J. Glass, S. Renals, P. Bell: "Current and Future Work: Arabic Speech Dialect Detection"
- V. Arora, A. Lahiri, H. Reetz: "Phonological Features for Automatic Speech Recognition"
- I. Casanueva: "Personalised dialogue management for users with speech disorders"
- S. Deena, M. Hasan, M. Doulaty, O. Saz, T. Hain: "Combining Feature and Model–Based Adaptation of RNNLMs for Multi–Genre Broadcast Speech Recognition"
- M. Doulaty, O. Saz, R. W. M. Ng, T. Hain: "Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation"
- J. Fainberg, P. Bell, S. Renals: "Utility of genre domains in the MGB corpus"
- G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, S. King: "Robust text–to–speech duration modelling using DNNs"
- L. Juarez Rivera, M. Russell, P. Jancovic: "Application of DNN–HMMs to children's speech recognition"
- B. Khaliq, O. Saz, T. Hain: "Segmentwise language model interpolation for lightly supervised alignment of broadcast subtitles"
- Y. Liu, C. Fox, M. Hasan, T. Hain: "The Sheffield Wargame Corpus – Day Two and Day Three"
- E. Loweimi, J. Barker, T. Hain: "Use of Generalised Nonlinearity in Vector Taylor Series Noise Compensation for Robust Speech Recognition"
- A. I. Masrani, Y. Gotoh: "Overlapped Interest and the Impact of Visual and Audio Information in the Human Perception"
- C. Pike, A. V. Beeston, T. Brookes, G. J. Brown, R. Mason: "Compensation for spectral and temporal envelope distortion caused by transmission channel" acoustics"
- A. Ragni, E. Dakin, X. Chen, M. J. F. Gales, K. M. Knill: "Multi–Language Neural Network Language Models"
- L. Rencker, W. Wang: "Sparsity based declipping of speech signals"
- M. Roddy, N. Harte: "Correlations between head movement and prosodic engagement features in dyadic conversations"
- P. Swietojanski, S. Renals: "LHUC and Differentiable Pooling for Acoustic Model Adaptation"
- D. Websdale, B. Milner: "A perceptually motivated loss function for DNN–based binary mask estimation for speech separation"
- J. H. M. Wong, M. J. F. Gales: "Hypothesis posterior student–teacher training"

Simulating the Visual Lombard Effect

Najwa Alghamdi, Steve Maddock, Guy J. Brown and Jon Barker

University of Sheffield, United Kingdom
{amalghamdi1, s.maddock, g.j.brown, j.p.barker}@sheffield.ac.uk

1. Abstract

Hearing-impaired individuals make significant use of facial signals during speech perception, however, they must also be able to deal with audio-only situations, a particular issue for the hard of hearing who use cochlear implants (CI) and hearing aids (HA). Previous work suggests that introducing visual speech in auditory training is effective at enhancing hearing abilities in subsequent audio-only situations ([1]). Our research investigates methods for increasing the effective of audio-visual training by using video processing techniques to enhance visual speech cues originating from the mouth area. One suggested enhancement method is to support the visual identification of phonetic information by automatically changing the speaking style of the speaker from normal to exaggerated. An example of exaggerated speech is Lombard speech which is produced in noisy conditions to support communication. Compared to speech produced in 'normal' conditions, Lombard speech is characterized by exaggerated audio and visual signals. For example, the audio signal shows increases in loudness and vowel duration, and the visual signal shows increases in the inner-lip area and the mouth and jaw opening ([2]). Our enhancement method simulates the Lombard visual signal for any `normal` audio signal, given the 'normal' visual signal.

To gain more understanding of the mouth behaviour in Lombard conditions in order to inform a simulation technique, we made audiovisual recordings of eight speakers reading digits in (i) normal (quiet) conditions & (ii) Lombard conditions using babble speech at 0 SNR in 65, 70 and 80 db SPL (figure 1). The technique used for the simulation is by using principle component analysis ([3, 4]) to approximate each video frame mouth landmarks (Lip_k) given the mean mouth shape $(mean(Lip))$ and a number of parameters, as $Lip_k = mean(Lip) + Pb$, where P is a set of eigenvectors of the covariance matrix of $(mean(Lip) - Lip_k)$; and b is the weight vector calculated using the formula $P^T(Lip_k - mean(Lip))$. Multiplying b with a scalar $\alpha > 1$ can account for lip shape exaggeration (figure 2). By comparing mean mouth width and height in the collected audiovisual Lombard recordings with different levels of automatic exaggeration generated from using different values of α , we noticed that $\alpha \in [1.2, 1.5]$ can generate a comparable simulation to the exaggeration level observed in Lombard speech at 80 db SPL. The next step in our research is to test the impact of using exaggerated videos in auditory training on improving CI-simulated speech intelligibility.

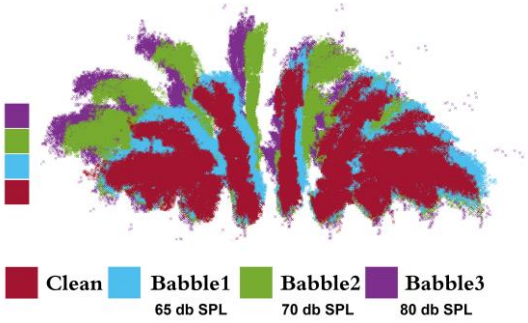


Figure 1: visualisation of mouth landmarks of the speakers tracked from the video recording frames (26 points per video frame) in normal and Lombard conditions. Landmarks points are normalized for visualization purpose.

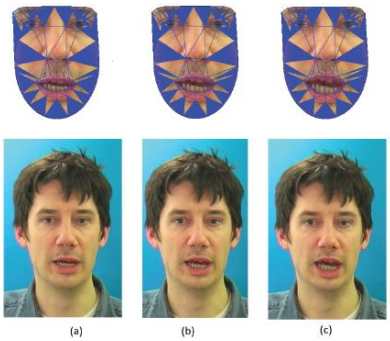


Figure 2: Video frames are re-animated by applying a 2D piecewise linear warping method using the estimated exaggerated mouth shapes. (a) The original frame; (b) and (c) are frames under two levels of exaggeration: α = 1.5 and 2.

2. References

[1] L. E. Bernstein, E. T. Auer Jr, S. P. Eberhardt, and J. Jiang, "Auditory perceptual learning for speech perception can be enhanced by audiovisual training," Frontiers in neuroscience, vol. 7, 2013.
[2] M. Fitzpatrick, J. Kim, and C. Davis. "Auditory and Auditory-Visual Lombard Speech Perception by Younger and Older Adults." Auditory-Visual Speech Processing (AVSP). 2013.
[3] B. Theobald, R. Harvey, S. Cox, G. Owen, and C. Lewis, "Lip-reading enhancement for law enforcement," in SPIE conference on Optics & Photonics for Counterterrorism and Crime Fighting, pp. 640 205–1, 2006.
[4] T. Cootes, E. Baldock, and J. Graham, "An introduction to active shape models," Image processing and analysis, pp.223–248, 2000.

Towards simple and robust audio-based detection of biomarkers for Alzheimer's disease

Sabah Al-Hameed [1], *Mohammed Benaissa* [1], *Heidi Christensen* [2]

[1] Department of Electronic and Electrical Engineering, University of Sheffield, United Kingdom
[2]Department of Computer Science, University of Sheffield, United Kingdom
{ssaal-hammed1, m.benaissa, heidi.christensen}@sheffield.ac.uk

1. Abstract

This project demonstrates the feasibility of using a simple and robust automatic method based solely on acoustic features to identify Alzheimer's disease (AD) with the objective of ultimately developing a low-cost home monitoring system for detecting early signs of AD. Different acoustic features, automatically extracted from speech recordings, are explored. Four different machine learning algorithms are used to calculate the classification accuracy between people with AD and a healthy control (HC) group. Feature selection and ranking is investigated resulting in increased accuracy and a decrease in the complexity of the method. Further improvements have been obtained by mitigating the effect of the background noise via pre-processing. Using DementiaBank data, we achieve a classification accuracy of 94.7%. This is an improvement on previous published results [1] whilst being solely audio-based and not requiring speech recognition for automatic transcription. The speech samples were obtained from the Dementiabank dataset [2].

The segments where the instructor speaks were removed using Praat software [3]. We also investigated a background noise reduction step in order to evaluate the performance of the proposed method under the effect of the background noise. A total of 263 acoustic features were extracted from the recordings in four different configurations. In the 1st configuration, all the 263 features were extracted and evaluated directly with the presence of the high background noise. The 2nd configuration the top (22) ranked features were selected from the 1st configuration. In the 3rd configuration, the spectral noise gating technique, using audacity [4] software was applied before extracting the features, while in the last configuration, only the top 20 ranked features were selected from the previous step to evaluate the accuracy. Weka software [5] was used for feature selection procedure and to train classifiers to discriminate between the two groups with k-fold = 10 as a cross validation. Fig. (2), showing the accuracies obtained and evaluated by the four classification algorithms. These results support our proposition for using only acoustic features for automatic detection and/or screening of AD at a low cost and within the home environment.

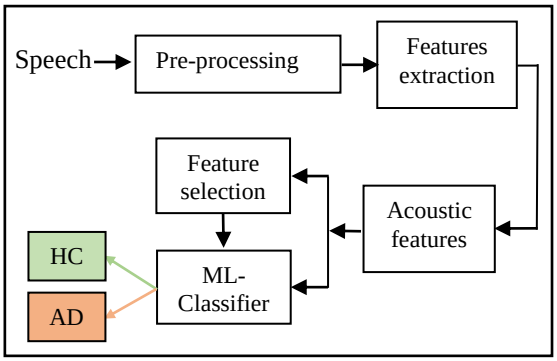


Figure. (1) Proposed method

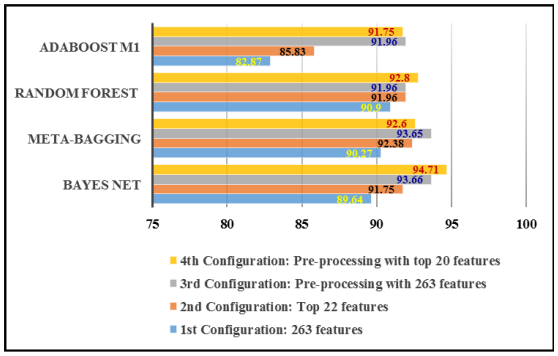


Figure (2) shows the performance under different running configurations

2. References

[1] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *J. Alzheimer's Dis.*, vol. 49, no. 2, pp. 407–422, 2015.
[2] "Dementia Bank." [Online]. Available: https://talkbank.org/DementiaBank/. [Accessed: 10-Dec-2015].
[3] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Available: http://www.fon.hum.uva.nl/praat/. [Accessed: 05-Jan-2016].
[4] "Audacity® " . Available: http://www.audacityteam.org/. [Accessed: 15-Jan-2016].
[5] "Weka 3: Data Mining Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 01-Feb-2016].

Current and Future Work: Arabic Speech Dialect Detection

Ahmed Ali[1], Sameer Khurana[1], Najeem Dehak[2], Patrick Cardinal[3], Shree harsha yella[4], James Glass[4], Steve Renals[5], Peter Bell[5]

[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
[2]JHU Center for Language and Speech Processing (CLSP), Baltimore, MD, USA
[3]École de technologie supérieure, Département de Ǵenie Logiciel et des TI, Montréal, Canada
[4]MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA
[5] Centre for Speech Technology Research, University of Edinburgh, UK
(skhurana,amali)@qf.org.qa, najim@jhu.edu

1. Abstract

The problem of Dialect Identification (DID) is a subset of a more general problem of Language Identification (LID). Whereas, LID is the process of classifying a speech segment into one of the many language classes, DID is the process of classifying speech segments into one of many dialects within the same language and hence is considered a more challenging problem than LID. Solving the problem of DID can help improve ASR systems by identifying dialectal segments from an untranscribed mixed-speech dataset. This process can help reduce the ASR word error rate (WER) for dialectal data by training ASR systems for each dialect, or by adapting the ASR models to a particular dialect.

Arabic Language is spoken in five main dialects: Egyptian (EGY), North African or Maghrebi (NOR), Gulf or Arabian Peninsula (GLF), Levantine (LAV), and Modern Standard Arabic (MSA). All dialects are historically related, but not synchronically, and are mutually unintelligible languages like English and Dutch.

Normal vernacular can be difficult to understand across different Arabic dialects. Arabic dialects are thus sufficiently distinctive, and it is reasonable to regard the DID task in Arabic as similar to the LID task in other languages.

In this presentation, we present our work on Arabic Dialect Detection of broadcast speech. We investigated three Vector Space Models (VSMs) of speech utterances.

- **Senone based Utterance VSM**: where each speech utterance is represented as a high dimensional sparse vector ($\boldsymbol{u}$): $\boldsymbol{u} = (A(f(u, s_1)), A(f(u, s_2)), \ldots, A(f(u, s_d)))$, where $f(u, s_i)$ is the number of times a senone s_i occurs in the speech utterance u, and A is the scaling function
- **Word based Utterance VSM**: similar to the senone based VSM, $\boldsymbol{u}$, in this case is: $\boldsymbol{u} = (A(f(u, w_1)), A(f(u, w_2)), ..., A(f(u, w'_d)))$, where $f(u, w_i)$ is the number of times a word w_i occurs in the speech utterance $\boldsymbol{u}$ and A is the scaling function
- **I-vector based utterance VSM**: 400 dimensional i-vectors were extracted for each speech utterance using a GMM-UBM, whose parameters were optimized to model the Bottleneck Features (BN) instead of the usual MFCC features

We did a thorough study of the usefulness of these three VSMs (and their combinations), in detecting the dialect of a speech utterance. We further discovered the huge difference in the DID performance of the classifiers in the in-domain and out-of-domain data. Extending on our previous work, we seek to improve the performance of our DID system on the out-of-domain data using semi-supervised learning and domain adaptation techniques. We discuss dialect identification errors in the context of dialect code-switching between Dialectal Arabic and MSA, and compare the error pattern between manually labeled data, and the output from our classifier. All the data used on our experiments have been released to the public as a language identification corpus.

We also present the future work and explain how semi-supervised learning techniques and domain adaptation methods can be leveraged to improve our dialect identification system.

Phonological Features for Automatic Speech Recognition

Vipul Arora[†]*, Aditi Lahiri*[†] *and Henning Reetz*[‡]

[†]Faculty of Linguistics, Philology and Phonetics, University of Oxford, United Kingdom
[‡]Goethe University, Frankfurt am Main, Germany

{vipul.arora, aditi.lahiri}@ling-phil.ox.ac.uk, reetz@em.uni-frankfurt.de

1. Abstract

Typical automatic speech recognition (ASR) systems use phones as the underlying representation of speech [1]. Acoustic models are trained to extract phones (i.e., their probabilities) from speech signals, that are then decoded into words and sentences with finite state transducers (FSTs). They handle the problem of speech variation with the help of statistical learning over large amounts of training data. The deep learning techniques have enabled them to effectively make use of larger amounts of data to learn the speech variations. Our research aims at enhancing the flexibility and controllability of ASR systems by using an alternate representation. We propose phonological features as the underlying representation of speech, instead of phones [2]. Phonological features will be advantageous in two major ways. Firstly, they provide natural classes to represent speech universally, enabling us to handle many speech variations within a language. They not only correspond well to the signal, but also model higher level speech variations in a robust and principled way. Thus, using them as underlying representation in ASR systems can simplify the learning problem, thereby cutting down the amount of training data and number of model parameters required. Secondly, they provide ways to easily transfer the knowledge across languages and dialects, thereby omitting the need to gather enormous data for retraining the acoustic model for the new scenario. Hence, they can be very useful for building ASR systems for under-resourced languages. The use of features instead of phones is also supported by our neuro-linguistic experiments [2], which show the universality of phonological features across different languages of the world.

Towards this end, we are first working towards detecting phonological features from speech signals. We define our features based on featurally underspecified lexicon model (FUL) [3]. Within FUL, an interesting property of phonological features is that they are underspecified. For instance, the place feature is specified as labial for the phone [m], but is unspecified for [n]. This allows [n] to easily change to [m] or [ng] in running speech, but leaves [m] unchanging. E.g., "green-bag" is often pronounced as "greem-bag" but [m] in "cream-desk" does not change to [n]. This property of the feature space makes the distance metric asymmetric. We have proposed a ternary-value representation to express the phonological feature vectors. Under this, each phone is characterised by a set of phonological features, each of which can take three values, viz., +1, 0 or -1. The value +1 denotes presence, -1 denotes absence and 0 leaves the feature unspecified. E.g., for [m], the feature LABIAL takes value +1, while for [n], LABIAL has value 0. The significance of unspecified value is that the detection or not-detection of that feature does not affect the decoding of the corresponding phone.

For this representation, a learning method has been developed to train a deep neural network (DNN) for detecting these features, while complying with the underspecification property. The network learns to detect a particular feature from the phones which have its value as +1 and to not-detect from phones having its value as -1. Its output is unconstrained for phones having its value as 0. For experiments, we have selected 18 features to characterise different phones in English language. The training and testing have been carried out over non-overlapping sets of speakers from TIMIT database. The obtained accuracies for feature detection are better than the conventional GMM based acoustic model adapted for this task.

The next step is to test the performance of these features for full-fledged ASR system for one language, i.e. English, and to compare the advantages of phonological features over conventional phone-based representation. Further, we would like to work towards making the feature detection system easily transferable or adaptable to another language. Our motivation for this undertaking comes, in principle, from phonological theory, and in practice, from the idea of zero-shot learning (ZSL) in the area of computer vision. In ZSL, a class is broken down into attributes (features) and the model is learned to detect these attributes; and a new (unseen) class is then detected based on the detection of (seen) attributes [4]. We would like to explore the paradigm of ZSL in area of speech so as to develop ways for cross-language model transfer for ASR.

2. References

[1] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[2] S. A. Cornell, A. Lahiri, and C. Eulitz, "What you encode is not necessarily what you store: Evidence for sparse feature representations from mismatch negativity," *Brain Research*, vol. 1394, pp. 79–89, 2011.

[3] A. Lahiri and H. Reetz, "Distinctive features: Phonological underspecification in representation and processing," *Journal of Phonetics*, vol. 38, no. 1, pp. 44–59, 2010.

[4] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Advances in Neural Information Processing Systems*, 2014, pp. 3464–3472.

Personalised dialogue management for users with speech disorders

Iñigo Casanueva

University of Sheffield, United Kingdom
i.casanueva@sheffield.ac.uk

1. Abstract

Voice controlled environmental control interfaces can be a life changing technology for users with disabilities. However, often these users suffer from speech disorders (e.g. dysarthria), making ASR very challenging. Acoustic model adaptation can improve the performance of the ASR, but the error rate will still be high for severe dysarthric speakers. POMDP-based dialogue management can improve the performance of these interfaces due to its robustness against high ASR error rates and its ability to find the optimal dialogue policy in each environment (e.g. the optimal policy depending on the dysarthria severity of the speaker or on the amount of acoustic data used to adapt the ASR). However, very little research has been done so far in dialogue model adaptation to unique users interacting with a system over a long period of time. This work shows how statistical dialogue management can be applied to an environmental control interface to improve its performance, explores different methods to adapt Gaussian process-based dialogue policies and RNN-based dialogue state trackers to unique users, and methods to adapt these models online as more data from the user becomes available.

Combining Feature and Model-Based Adaptation of RNNLMs for Multi-Genre Broadcast Speech Recognition

Salil Deena, Madina Hasan, Mortaza Doulaty, Oscar Saz and Thomas Hain

Speech and Hearing Research Group, The University of Sheffield, UK

{s.deena, m.hasan, m.doulaty, o.saz, t.hain}@sheffield.ac.uk

1. Abstract

Language models (LMs) play a key role in modern ASR and machine translation systems as they ensure that the output respects the pattern of the language in question. n-gram LMs dominated ASR for decades until RNNLMs [1] were introduced and found to give significant gains in performance. Recurrent neural network language models (RNNLMs) have consistently outperformed n-gram language models when used in automatic speech recognition (ASR). This is because RNNLMs provide robust parameter estimation through the use of a continuous-space representation of words, and can generally model longer context dependencies than n-grams. It is found that n-gram LM and RNNLM contributions are complementary and state-of-the-art ASR systems involve interpolation between the two types of models.

In automatic speech recognition, language context is generally heavily influenced by the domain, which can include topic, genre and speaking style. RNNLMs trained on a text corpus provide an implicit modelling of such contextual factors. However, it has been found that domain adaptation of RNNLMs to small amounts of matched in-domain text data provide significant improvements in both perplexity (PPL) and word error rate (WER) [2, 3, 4]. RNNLM adaptation can be categorised as either feature-based [3, 4] or model-based [2, 5, 6]. The former involves augmenting the input to the RNNLM with auxiliary features that encode domain information whilst the latter involves adapting the network to the new domain. Model-based RNNLM adaptation can either involve fine-tuning, which involves further training the RNNLM with matched in-domain data or the introduction of adaptation layer(s) to adapt the network to new domains.

Whilst feature-based RNNLM adaptation was shown to outperform domain fine tuning [4], it is required that the auxiliary features be known at the time of model training and thus can be inflexible, as it requires for the whole model to be re-trained should altered features become available. Domain fine-tuning somehow addresses such limitation as the RNNLM can be fine-tuned using newly available domain-specific data and do not require retraining of the whole RNNLM. However, the shared information between domains is not properly modelled. A combination of feature and model adaptation can thus provide the best solution in many instances. We provide a detailed comparison of both types of adaptation on RNNLMs trained on both small and large text corpora and propose novel techniques for RNNLM adaptation, including the linear hidden network (LHN) [7] adaptation layer as well as hybrid adaptation methods, and are evaluated on a broadcast media transcription task [8]. The gains obtained with RNNLM adaptation on a system trained on 700h. of speech are consistent using both RNNLMs trained on a small (10M words) and large set (660M words), with 10% perplexity and 2% word error rate improvements on a 28.3h. test set.

2. References

[1] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH'10: Proc. of the 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 1045–1048.

[2] J. Park, X. Liu, M. J. F. Gales, and P. C. Woodland, "Improved neural network based language modelling and adaptation," in *INTERSPEECH'10: Proc. of the 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 1041–1044.

[3] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model." in *SLT'12: Proc. of the IEEE workshop on Spoken Language Technologies*, 2012, pp. 234–239.

[4] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M. J. F. Gales, and P. C. Woodland, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition," in *INTERSPEECH'15: Proc. of the 16th Annual Conference of the International Speech Communication Association*, 2015, pp. 3511–3515.

[5] T. Alumäe, "Multi-domain neural network language model," in *INTERSPEECH'13, 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 2182–2186.

[6] O. Tilk and T. Alumäe, "Multi-domain recurrent neural network language model for medical speech recognition," in *Human Language Technologies - The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014, Kaunas, Lithuania, September 26-27, 2014*, 2014, pp. 149–152.

[7] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. de Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models." *Speech Communication*, vol. 49, no. 10-11, pp. 827–835, 2007.

[8] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster, and P. Woodland, "The MGB challenge: Evaluating multi–genre broadcast media transcription," in *ASRU'15: Proc. of IEEE workshop on Automatic Speech Recognition and Understanding*, Scottsdale, AZ, 2015.

Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation

Mortaza Doulaty, Oscar Saz, Raymond W. M. Ng, Thomas Hain

Speech and Hearing Group, Department of Computer Science, University of Sheffield, UK

{mortaza.doulaty, o.saztorralba, wm.ng, t.hain}@sheffield.ac.uk

1. Abstract

This paper presents a new method for the discovery of latent domains in diverse speech data, for the use of adaptation of Deep Neural Networks (DNNs) for Automatic Speech Recognition. Our work focuses on transcription of multi-genre broadcast media, which is often only categorised broadly in terms of high level genres such as sports, news, documentary, etc. However, in terms of acoustic modelling these categories are coarse. Instead, it is expected that a mixture of latent domains can better represent the complex and diverse behaviours within a TV show, and therefore lead to better and more robust performance. We propose a new method, whereby these latent domains are discovered with Latent Dirichlet Allocation (LDA) [1], in an unsupervised manner.

LDA was originally used for topic modelling of text corpora; however, it is a generic model and can be applied to other tasks, such as object categorisation and localisation in image processing [2], automatic harmonic analysis in music processing [3], acoustic information retrieval in unstructured audio analysis [4] and our previous work for domain adaptation of GMM/HMM systems [5].

LDA domain posteriors are used to adapt DNNs using the Unique Binary Code (UBIC) representation for the LDA domains. TV broadcasts from the BBC were selected for the experiments. The data is identical to the one defined and provided for the 2015 Multi–Genre Broadcast (MGB) Challenge [6]. The shows were in 8 genres: advice, children's, comedy, competition, documentary, drama, events and news. Acoustic model training data was around 2,000 shows and the development data for the task was 47 shows.

Table 1 presents the WER of baseline and adapted models for all of the eight genres. Latent Domain adaptive Training (LDaT) reduces the WER from 33.3% to 30.6%, which is even better than speaker adapted DNN (31.4%). Combining speaker adaptation and domain adaptation (SAT+LDaT, linear input transformation for the speaker and bias adaptation for the latent domain) yields 28.9%, which is 13% relative WER reduction compared to the baseline DNN model and 8% relative improvement over the speaker adapted DNN. This also suggests that LDA inferred domains were not speaker clusters (since combining two adaptations still improves the performance). Because of the diverse nature of the data used, WER differs a lot across genres. Namely comedy and drama had the highest errors (43.8% and 45.0% respectively with LDaT+SAT models) showing the difficult nature of these genres. On the other hand, news had the lowest WER (14.3%). The WER diversity across the genres was consistent between all of the four models presented in table 1.

The proposed method lends itself to several future investigations. In the current LDA domain representation, each domain is described as a point on one of the axes of a high–dimensional space, where all have same distance from each other. Representing these points differently so that similar domains became closer in that space and verifying how that improves the performance can be an interesting problem to verify as a future work.

Table 1: Per–genre WER for all of the models

Adaptation	WER (%)								
	Advice	Child.	Comedy	Compet.	Docum.	Drama	Even.	News	**Overall**
–	27.6	29.1	47.8	28.2	31.3	52.0	38.1	17.9	**33.3**
SAT	26.2	27.5	46.1	25.9	29.8	49.3	35.8	15.9	**31.4**
LDaT	25.8	27.8	45.1	25.7	28.9	47.7	33.5	15.7	**30.6**
LDaT+SAT	24.2	26.5	43.8	23.6	27.3	45.0	31.6	14.3	**28.9**

2. References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. of ICCV*, Beijing, China, 2005.

[3] D. Hu and L. K. Saul, "A probabilistic topic model for unsupervised learning of musical key-profiles." in *Proc. of ISMIR*, Kobe, Japan, 2009.

[4] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," in *Proc. of WASPAA*, New Paltz NY, USA, 2009, pp. 37–40.

[5] M. Doulaty, O. Saz, and T. Hain, "Unsupervised domain discovery using latent dirichlet allocation for acoustic modelling in speech recognition," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[6] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster, and P. Woodland, "The MGB Challenge: Evaluating multi-genre broadcast media recognition," in *Proc. of ASRU*, Arizona, USA, 2015.

Utility of genre domains in the MGB corpus

Joachim Fainberg, Peter Bell, Steve Renals

University of Edinburgh, United Kingdom

j.fainberg@sms.ed.ac.uk, peter.bell@ed.ac.uk, s.renals@ed.ac.uk

1. Abstract

We present ongoing work on the MGB corpus [1]. Each show in the corpus is annotated with its corresponding genre. However, initial experiments suggest that factorising the data by genre reduces performance (Figure 1). The annotated genres may not correspond to acoustic domains. To test this notion, we first extract i-vectors [2] at the show level and compare pairwise likelihoods given a Probabilistic Linear Discriminant Analysis (PLDA) [3] model. Results suggest that the series from which a show originates yields a much stronger underlying acoustic space than genres (Table 1), though we also observe higher average likelihoods within genres than across (1.16 and -0.92, respectively). This suggests that the genres do convey some information that it might be useful to employ. A preliminary experiment with auxiliary features, encoding the genre labels as 1-hot vectors, yielded 0.9% absolute reduction in Word Error Rate. An experiment with genre labels as an additional task during training in a multi-task setup did, however, not yield improvements. We are investigating methods of incorporating and making best use of these, and other labels. In particular factorised methods, such as factorising i-vectors [4] or bottleneck layers [5].

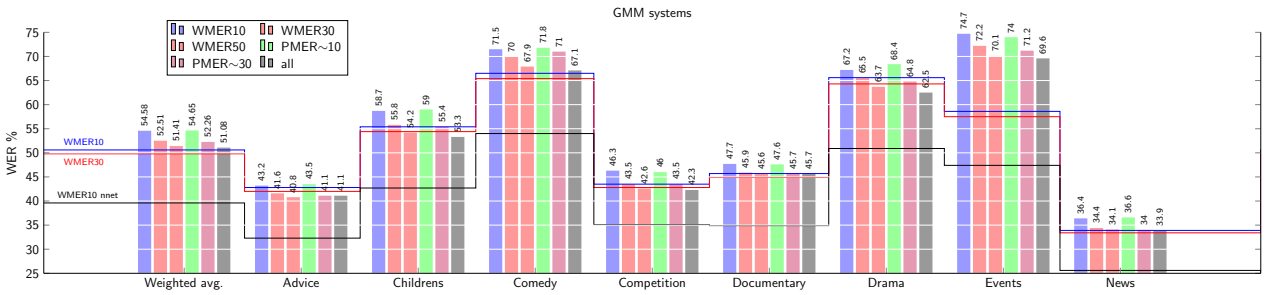


Figure 1: *Baseline average and genre-dependent models*

Table 1: *Average likelihoods for show i-vectors. Mean genre excludes comparisons within series.*

	Advice	Childrens	Comedy	Comp.	Docu.	Drama	Events	News
Mean genre	1.37	0.36	2.24	-1.21	1.02	2.15	-0.21	2.50
Mean series	6.09	6.46	7.17	11.27	7.87	6.49	6.59	5.03

2. References

[1] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster *et al.*, "The MGB Challenge: Evaluating multi-genre broadcast media transcription," in *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding, Scottsdale, AZ*, 2015.

[2] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 55–59.

[3] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, 2007, pp. 1–8.

[4] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc Interspeech*, 2014.

[5] M. Ferras and H. Bourlard, "MLP-based factor analysis for tandem speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6719–6723.

Robust text-to-speech duration modelling using DNNs

Gustav Eje Henter, Srikanth Ronanki, Oliver Watts, Mirjam Wester, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK

gustav.henter@ed.ac.uk

1. Abstract

Accurate modelling and prediction of speech-sound durations is an important component in generating more natural synthetic speech. Deep neural networks (DNNs) offer a powerful modelling paradigm, and large, found corpora of natural and prosodically-rich speech are easy to acquire for training DNN models. Unfortunately, poor quality control (e.g., transcription errors) as well hard-to-predict phenomena such as reductions and filled pauses are likely to complicate duration modelling from found data. To mitigate issues caused by these idiosyncrasies, we propose to improve modelling and prediction of speech durations using methods from *robust statistics*. These are able to disregard ill-fitting points in the training material – errors or other outliers – in order to describe the typical case better. For instance, parameter estimation can be made robust by changing from maximum likelihood estimation (MLE) to a robust fitting criterion based on the density power divergence (a.k.a. the β-divergence) [1, 2]. Alternatively, the standard approximations for output generation with multi-component mixture density networks (MDNs) [3] can be seen as a heuristic for robust output generation.

To evaluate the potential benefits of robust techniques, we used 175 minutes of found data from a free audiobook to build several text-to-speech (TTS) systems, described in Table 1, with either conventional or robust DNN-based duration prediction. The objective results, in Figure 1, indicate that robust methods described typical speech durations better than the baselines. (Atypical, poorly predicted durations may be due to transcription errors, known to exist also in the test data, that make some FRC durations unreliable.) Similarly, subjective evaluation using a hybrid MUSHRA/preference test with 21 listeners, each scoring 18 sets of same-sentence stimuli, found that listeners significantly preferred synthetic speech generated using robust methods over the baselines, as shown in Figure 2.

Label	Duration prediction method	Robust?	Label	Duration prediction method	Robust?
VOC	Vocoded speech (top line waveform)	-	MLE1	Gaussian MLE-fitted DNN (baseline)	no
FRC	Oracle durations from forced alignment against held-out speech	-	MLE3	3-component deep Gaussian MDN only synthesising from the heaviest component	yes
BOT	Monophone mean duration (bottom line)	no	B75	Gaussian DNNs fit using β-divergence, tuned to ignore ≈ 25 or 50% of datapoints	yes
MSE	Minimum mean-square error (baseline)	no	B50		yes

Table 1: *TTS systems in evaluation. Except for vocoded speech, all used the same DNN acoustic model but different duration predictors.*

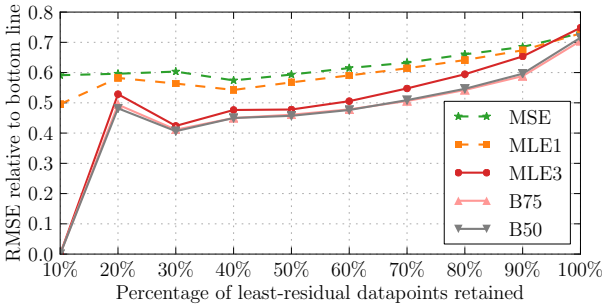


Figure 1: *Relative RMSE (frames per phone) between predicted and forced-aligned (FRC) durations on progressively larger and less well explained test-data subsets. Performance is normalised to place BOT at 1.0. Robust systems (solid) outperform non-robust baselines (dashed) on the majority of datapoints.*

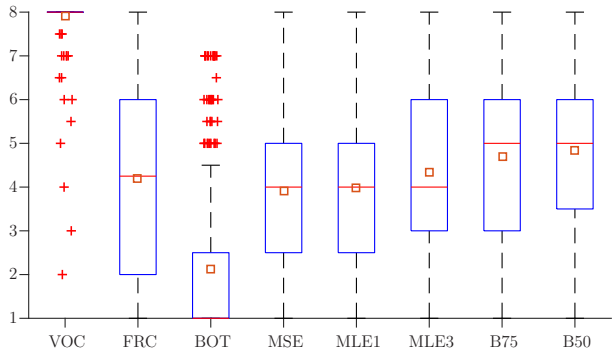


Figure 2: *Aggregated ranks (higher is better) from listening test. Red lines are medians, orange squares means; box edges are at 25 and 75% quantiles. Again, robust methods trump baselines.*

2. References

[1] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[2] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep. Research Memo 802, June 2001. [Online]. Available: http://www.ism.ac.jp/~eguchi/pdf/Robustify_MLE.pdf

[3] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, 2014, pp. 3844–3848.

Application of DNN-HMMs to children's speech recognition

Luciano Juarez Rivera, Martin Russell, Peter Jancovic

School of Engineering, University of Birmingham
{lej497, m.j.russell, p.jancovic}@bham.ac.uk

Abstract

During recent years, technological advances have allowed a much deeper exploration in the field of Automatic Speech Recognition (ASR) as the computational power of machines has increased. As an effect, Hidden Markov Model – Deep Neural Network hybrid systems (HMM-DNN) are becoming a widely used modelling technique for ASR as it has been demonstrated that they can outperform the traditional HMM-GMM approach [1]. However, the major benefit from these advances has been for ASR tasks on adult speech. It is well known that ASR is significantly more difficult for child speech than for adult speech [6], and that acoustic variability is much greater for children's speech [4]. Some of these differences are due to physiological factors, such as children's shorter vocal tracts and smaller larynxes, while increased variability may be due to developing motor control skills or cognitive factors associated with language acquisition. Although some effort has been applied to the application of HMM-DNN systems to children's speech [3, 5] this area is still relatively unexplored. Additionally, various normalization techniques to support acoustic modelling in ASR have been developed [2, 5], but there is still not a universal recipe to be implemented with confidence in the development of ASR systems for children's speech. In this work, we are exploring the application of HMM-DNN systems to ASR, methods for improving its performance and alternatives to enhance children's ASR systems to be possibly used in important applications such as speech therapy assistants.

References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 30–42, 2012.

[2] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children.," in Interspeech, pp. 2749–2752, 2005.

[3] A, Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners". In Interspeech, pp. 1468-1472, 2014.

[4] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on, pp. 22–25, IEEE, 2007.

[5] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in Spoken Language Technology Workshop (SLT), 2014 IEEE, pp. 135–140, IEEE, 2014.

[6] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 1, pp. 349–352, IEEE, 1996.

Segmentwise language model interpolation for lightly supervised alignment of broadcast subtitles

Bilal Khaliq, Oscar Saz, Thomas Hain

Speech and Hearing Research Group, The University of Sheffield, UK

{b.khaliq,o.saztorralba,t.hain}@sheffield.ac.uk

1. Abstract

Alignment of subtitles is a new task related to speech transcription which requires to identify words from a given text that are spoken in the recording, and to provide their timing. The task was defined in the context of broadcast media processing, and evaluated on the context of the Multi-genre Broadcast (MGB) challenge, however it has also relevance to semi-supervised training of acoustic models. Typical alignment systems perform speech recognition with Languages Models (LMs) that are biased towards the target material, on a per show basis. In this paper, a finer level of interpolation is presented. Interpolation mixture weights are inferred at segment level thus indicate the degree of reliance on the subtitle text. Making use of initial decoding output, classifiers are constructed to learn the LM combination that minimises the error for each segment. With the LM interpolation applied individually for each segment, significant gains in lightly supervised Automatic Speech Recognition (ASR) performance are obtained.

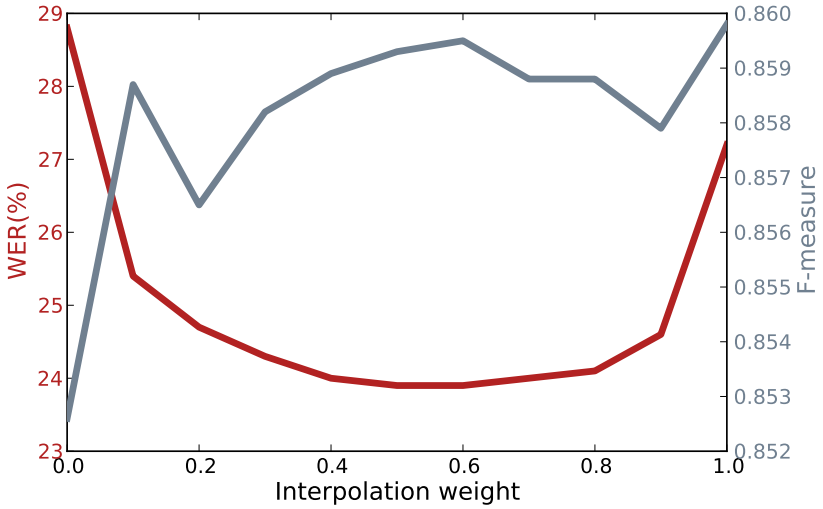


Figure 1: *Effect of the interpolation weights on WER and F–measure results on MGB development data.*

Table 1: *Lightly supervised decoding and lightly supervised alignment results.*

Baseline system (interpolation weight=0.5)				
Dataset	WER	Precision	Recall	F–measure
MGB Development	**23.9%**	0.8459	0.8731	**0.8593**
MGB Evaluation	**24.8%**	0.8174	0.8697	**0.8427**
Oracle system (best interpolation weights)				
Dataset	WER	Precision	Recall	F–measure
MGB Development	**21.1%**	0.8456	0.8771	**0.8611**
Proposed system (estimated interpolation weights)				
Dataset	WER	Precision	Recall	F–measure
MGB Development	**22.7%**	0.8439	0.8747	**0.8590**
MGB Evaluation	**23.9%**	0.8179	0.8697	**0.8430**

The Sheffield Wargame Corpus - Day Two and Day Three

Yulan Liu[1†], *Charles Fox*[2], *Madina Hasan*[1], *Thomas Hain*[2]

[1]The University of Sheffield, United Kingdom
[2] The University of Leeds, United Kingdom
[†]acp12yl@sheffield.ac.uk

1. Abstract

Improving the performance of distant speech recognition is of considerable current interest, driven by a desire to bring speech recognition into peoples homes. Standard approaches to this task aim to enhance the signal prior to recognition, typically using beamforming techniques on multiple channels. Only few real-world recordings are available that allow experimentation with such techniques. This has become even more pertinent with recent works with deep neural networks aiming to learn beamforming from data. Such approaches require large multi-channel training sets, ideally with location annotation for moving speakers, which is scarce in existing corpora. This poster presents a freely available and new extended corpus of English speech recordings in a natural setting, with moving speakers. The data is recorded with diverse microphone arrays, and uniquely, with ground truth location tracking. It extends the 8.0 hour Sheffield Wargames Corpus released in Interspeech 2013, with a further 16.6 hours of fully annotated data, including 6.1 hours of female speech to improve gender bias. Additional blog-based language model data is provided alongside, as well as a Kaldi baseline system. Results are reported with a standard Kaldi configuration, and a baseline meeting recognition system.

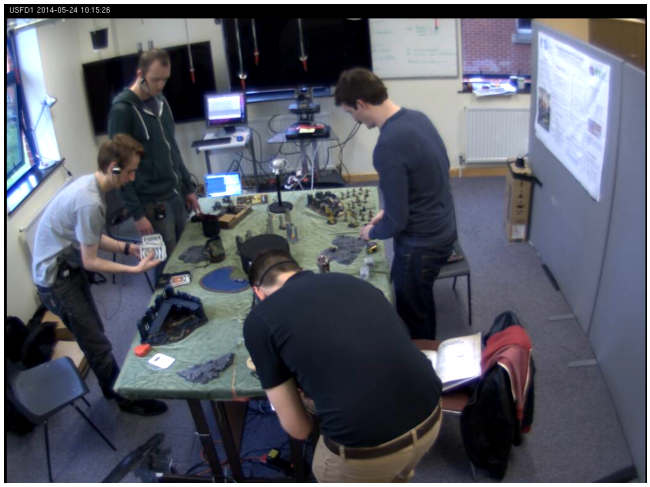


Table 1: *SWC corpora statistics.*

	SWC1	SWC2	SWC3	overall
#session	10	8	6	24
#game	4	4	3	11
#unique speaker	9	11	8	22
gender	M	M	F&M	F&M
#unique mic	96	71	24	103
#shared mic	-	-	-	24
speech duration	8.0h	10.5h	6.1h	24.6h
#speech utt.	14.0k	15.4k	10.2k	39.6k
duration per utt.	2.1s	2.5s	2.2s	2.2s
#word per utt.	6.6	7.9	5.5	6.8
vocabulary	4.4k	5.7k	2.9k	8.5k
video	√	√	-	√
location	√	√	√	√

Use of Generalised Nonlinearity in Vector Taylor Series Noise Compensation for Robust Speech Recognition

Erfan Loweimi, Jon Barker and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK
{eloweimi1, j.p.barker, t.hain}@sheffield.ac.uk

1. Abstract

Vector Taylor Series compensation is among the most effective methods for enhancing the robustness of the automatic speech recognition systems against noise. However, its formulation relies on the additive property of log-domain filterbank energies. This means that although it can be directly employed with MFCCs it cannot be used with features that use power transformation, i.e. generalized logarithmic function (GLF) instead of logarithm, e.g., PLP, PNCC and various phase-based features. Replacing the logarithm with the GLF, turns filterbank additions into multiplications and complicates the VTS linearisation process.

There are strong motivations for producing a VTS formulation that can be applied with GLF-based features. The GLF has an advantage that it has one degree of freedom and when this parameter tends to zero, this function acts similar to logarithm function. This transform in statistics literature is known as Box-Cox transformation (BCT) and has a remarkable effect on the statistical behaviour of the data. Arguably it is claimed that this transformation under using optimal parameter can enhance the linearity, normality and decreases Heteroscedasticity (variance stabilization). Among these, normality is of particular interest due to two reason. First, the model which is usually used for clean data during VTS noise compensation process is GMM. Although it can fit most of the pdfs under availability of enough data and sufficient components, enhancing the normality of data results in needing less Gaussians and as such less data for achieving a reasonable fit. Second, if in the back-end, GMM is used along with HMM, having a data with a better normality is advantageous and lead to a better fit.

As such, having one degree of freedom which could be fine-tuned for each data or task, together with noteworthy statistical influence on the pdf of the features make extending the VTS to the features utilize power transformation a worthwhile attempt. We refer to this transformation Generalized non-linearity (GN) as in speech community (1984) it is known as GLF and in statistics (1964) is known as BCT.

Out of these points, in this poster, a set of novel formulation for the VTS assuming that GN is used instead of logarithmic function is derived in both frequency and cepstrum domains. The efficacy is also put into test in Aurora4 databases. This formulation expands the applicability of the VTS method and results in absolute 12.2% and 2.0% WER reduction compared with MFCC and conventional VTS, respectively (Table 1). As such this approach smooths the way to extend VTS framework to generalized-MFCC, PLP, PNCC and phase-based features and consequently opens up a broad avenue for future researches.

Table 1: *Word error rates (WER) for Aurora-4 (Ave=* $\frac{A+6B+C+6D}{14}$).

	Test Set A	Test Set B	Test Set C	Test Set D	Ave
MFCC	6.8	33.4	23.8	50.2	38.0
VTS	6.6	22.0	22.7	37.9	27.8
gVTS	6.8	19.5	21.5	36.1	25.8

Overlapped Interest and the Impact of Visual and Audio Information in the Human Perception

Atiqah Izzati Masrani, Yoshihiko Gotoh

University of Sheffield, United Kingdom

amasrani1@sheffield.ac.uk, y.gotoh@sheffield.ac.uk

1. Abstract

When presented with multiple videos, how do one select the video that fits one's requirements the most? Video descriptions is one way of addressing this issue. However, manual annotations (annotations that are written by human) can be both cost and time consuming. Therefore automatic description generation for video data has been an active research field. Textual representations are often opted compared to graphical representations because it has faster retrieval time and less space consuming. It is commonly represented in the form of keywords or natural language. Although keywords work best for faster video categorization and retrieval, natural language provides a more understandable and less ambiguous description of the video [1].

Recent works of natural language generations for video optimizes the visual information [2] [3]. We have conducted an experiment to investigate the impact of visual and audio information on how the participants annotate a video. The participants were first asked to watch the video without the audio on and write a brief description of the video. Afterwards, the participants were asked to watch the same video, this time with the audio on and again write a brief description of the video. Once the hand annotations have been compiled and preprocessed, we calculated the tf:idf values for every classes. We also used the Jaccard Coefficient [4] to measure the similarity values among the annotations in the same class. Based on our findings, we discovered that the hand annotations without audio has a higher average of term relevance (1) and similarity value (2) throughout the video corpora.

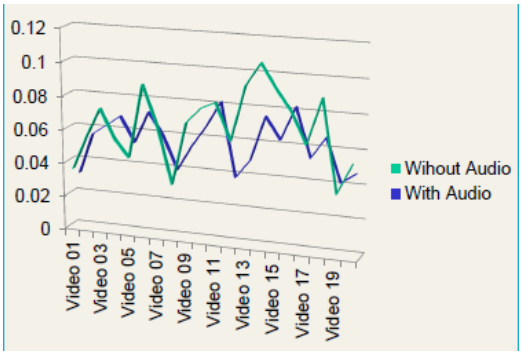


Figure 1: *Average* tf:idf *values for the hand annotations.*

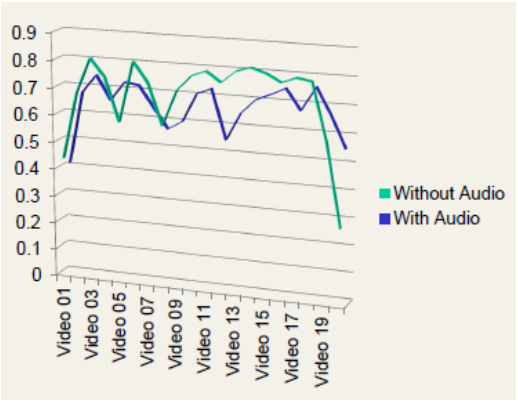


Figure 2: *Average similarity values for the hand annotations.*

However, there are two classes (Video ID 05 and 19) where the hand annotation with audio have higher average of term relevance and similarity value as compared to the hand annotation without audio. Our hypothesis is that these videos (Video ID 05 and 19) pose interesting scenes that intrigue the participants to describe the events in detail. As opposed to the other videos, the participants tend to describe what is visually present and were selective to include audio information. This causes the diversity in the hand annotations with audio. We aim to further investigate what set these videos apart from the others and how audio information increases the term relevance and similarity in the descriptions. This will be used as our basis for modeling the methodology to generate natural language descriptions for video data incorporating both its visual and audio information.

2. References

[1] A. I. Masrani and Y. Gotoh, "Corpus generation and analysis: Incorporating audio data towards curbing missing information," in *Proceedings of the 1st International Workshop on Knowledge Discovery on the WEB*, 2015.

[2] M. Khan, N. AlHarbi, and Y. Gotoh, "A framework for creating natural language descriptions of video streams," *Information Sciences*, vol. 303, p. 6182, 2015.

[3] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vision*, vol. 50, pp. 171–184, 2002.

[4] A. Huang, "Similarity measures for text document clustering," 2008.

Compensation for spectral and temporal envelope distortion caused by transmission channel acoustics

Cleo Pike[1,2], Amy V Beeston[3], Tim Brookes[2], Guy J Brown[3], and Russell Mason[2]

[1] School of Psychology and Neuroscience, University of St Andrews,
[2] Institute of Sound Recording, University of Surrey,
[3] Department of Computer Science, University of Sheffield
c.pike@surrey.ac.uk, a.beeston@sheffield.ac.uk

1. Abstract

Recognition of a sound's identity depends on accurate perception of its spectral and temporal envelopes, both of which are distorted in everyday listening spaces by a multitude of spectrally-altered and temporally-delayed reflections of the original sound. For example, room reflections might distort the frequency regions in which resonance cues signal the formant /e/, such that an /i/ is recognised instead (Watkins 1991, Pike et al. 2014). Similarly, room reflections might obscure dips in the temporal envelope which would otherwise cue the presence of an unvoiced plosive such as /t/ (Watkins 2005, Beeston et al. 2014). However, despite these distortions to the signal, human recognition of a sound's identity remains remarkably robust in most rooms.

A number of auditory mechanisms are thought to reduce the perceptual effects or confusions arising from acoustic distortions caused by room reflections. Our poster sets out a skeleton conceptual model of the compensation process for both spectral envelope and temporal envelope distortion, illustrating similarities between the two compensatory processes. The model collates findings from previous listening experiments both by the authors and by other researchers, and identifies 'blind spots' where insufficient data exists at present. This leads us to propose and discuss further behavioural, neuroscientific and computational experiments designed to test for common neural and psychological mechanisms behind compensation for spectral and temporal envelope distortion.
.

2. References

[1] Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. J Acoust Soc Am, 90(6), 2942-2955.
[2] Pike, C., Brookes, T., and Mason, R. (2014). Auditory compensation for spectral colouration, 137th Convention of the Audio Engineering Society, Los Angeles, USA. 9-12 Oct, preprint 9138
[3] Watkins, A. J. (2005). Perceptual compensation for effects of reverberation in speech identification. J Acoust Soc Am, 118(1), 249–262.
[4] Beeston, A. V., Brown, G. J. and Watkins, A. J. (2014). Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli. J Acoust Soc Am, 136(6), 3072–3084.

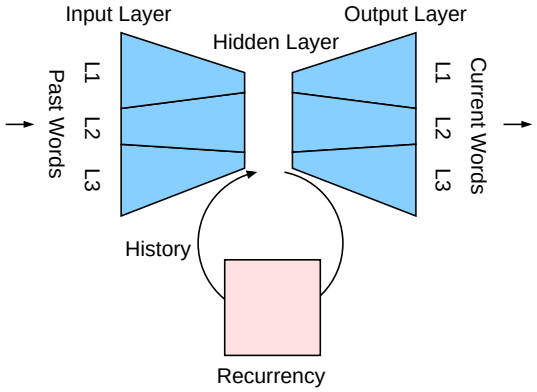
Multi-Language Neural Network Language Models

Anton Ragni, Edgar Dakin, Xie Chen, Mark J. F. Gales, Kate M. Knill

University of Cambridge, United Kingdom
{ar527,ed408,xc257,mjfg,kate.knill}@eng.cam.ac.uk

1. Abstract

In recent years there has been considerable interest in neural network based language models. These models typically consist of vocabulary dependent input and output layers and one, or more, hidden layers. A standard problem with these networks is that large quantities of training data are needed to robustly estimate the model parameters. This poses a challenge when only limited data is available for the target language. Motivated by the success of multilingual training of deep neural network in acoustic model, the multilingual training of recurrent neural network language model is investigated in this poster. A general solution that allows data from any language to be used is proposed. Here, only the input and output layers are language dependent whilst hidden layers are shared, language independent, which is illustrated as Figure 1. This multi-task training set-up allows the quantity of data available to train the hidden layers to be increased. This multi-language network can be used in a range of configurations, including as initialisation for previously unseen languages. As a proof of concept we examines multilingual recurrent neural network language models. Experiments are conducted using language packs released within the IARPA Babel program. A total of 14 diverse languages provided by the IARPA Babel program were considered. Results suggest that shared hidden layer representations can consistently help to reduce perplexity of each individual languages. Small but consistent improvement in terms of WER is also obtained.



2. References

[1] A. Ragni, E. Dakin, X. Chen, M. J. F. Gales and K. M. Knill “Multi-Language Neural Network Language Models,” *accepted by Interspeech*, 2016.

Sparsity based declipping of speech signals

Lucas Rencker, Wenwu Wang

CVSSP, University of Surrey

Abstract

Speech signals are often subject to non-linear distortion such as clipping. Clipping usually occurs when the energy of the signal exceeds the dynamic range limitations of the recording device, and the waveform is truncated above a certain threshold. This results in perceptuals artifacts due to the introduction of unwanted harmonics, but some recent studies have also shown that clipping strongly affects Automatic Speech Recognition rates [1]. Sparsity-based declipping techniques recently attracted some interest. Treating the clipped samples as missing samples, the declipping problem has been treated as an *Audio Inpainting* problem [2]. The goal is then to estimate the sparse representation of the speech signal in some overcomplete dictionary of atoms, based on only on the reliable (i.e. non-clipped) samples. In the clipping scenario, we can also add some knowledge on the amplitude of the reconstructed samples: we know that the reconstructed samples must lie above the clipping level. The declipping problem can then be formulated as a optimization problem with sparsity and amplitude constraints. This problem has been tackled in [2] using an Orthogonal Matching Pursuit (OMP) algorithm to first estimate a sparse support of non-zero atoms, followed by an amplitude-constrained least-squares. More recently, the Alternating Direction Method of Multipliers (ADMM) has been adapted to the declipping problem [3]. The ADMM allows to iteratively minimize an ℓ_0 norm and a constrained least squares. Although this algorithm demonstrated very good performance in terms of SNR improvement, experiments showed that the algorithm did not always converge. This is mainly due to the non-convexity of the ℓ_0 norm. In this paper, we propose to reformulate the problem proposed in [3], by relaxing the ℓ_0 norm into an ℓ_1 norm. The proposed formulation can then be seen as a constrained Lasso, and solved in a similar way using ADMM.

Correlations between head movement and prosodic engagement features in dyadic conversations

Matthew Roddy, Naomi Harte

Trinity College Dublin, Ireland

roddym@tcd.ie, nharte@tcd.ie

1. Abstract

Training audio-visual models of conversational engagement relies on building datasets that are generally hand-labeled for engagement. The task of producing these datasets is laborious and the datasets have the potential to be unreliable (or limited in their utility) if they are labeled by non-experts. We seek to find an alternative approach that is based on finding correlates in engagement features across modalities. Previous research has proposed that visual features derived from a person's movement are correlated with prosodic indicators of engagement during monologues. In conversational settings, the gestural behaviour of speakers also serves as a method of communicating listener attentiveness through the use of backchannels. We develop a method of extracting visual features that incorporates information from backchannels and that increases the correlation between audio and visual features.We also find that the level to which these movement features were correlated with prosodic features was dependent on the individual speaker's characteristic behaviour.

LHUC and Differentiable Pooling for Acoustic Model Adaptation

Pawel Swietojanski, Steve Renals

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

p.swietojanski@ed.ac.uk, s.renals@ed.ac.uk

1. Abstract

We present a deep neural network (DNN) acoustic model adaptation techniques based on learning hidden unit contributions (LHUC) and parametrised and differentiable pooling operators. Unsupervised acoustic model adaptation is cast as the problem of updating the decision boundaries implemented by each LHUC scaler or pooling operator. In particular, we experiment with two types of pooling parametrisations: learned L_p-norm pooling and weighted Gaussian pooling, in which the weights of both operators are treated as speaker-dependent. We perform investigations using three different large vocabulary speech recognition corpora: AMI meetings, TED talks and Switchboard conversational telephone speech. We demonstrate that both and differentiable pooling operators provide a robust and relatively low-dimensional way to adapt acoustic models, with relative word error rates reductions ranging from 5–20% with respect to unadapted systems. Differentiable pooling versions are themselves are better than the baseline fully-connected DNN-based acoustic models This presentation is an extract from our two published works [1] and [2].

2. Motivations, Methods, Findings and Conclusions

Deep neural network (DNN) acoustic models have significantly extended the state-of-the-art in speech recognition and are known to be able to learn significant invariances through many layers of non-linear transformations [3]. If the training and deployment conditions of the acoustic model are mismatched then the runtime data distribution can differ from the training distribution, bringing a degradation in accuracy, which may be addressed through explicit adaptation to the test conditions. In fact, it has been experimentally demonstrated that the invariance of the internal representations with respect to variabilities in the input space increases with depth (the number of layers) and that the DNN can interpolate well around training samples but fails to extrapolate if the data mismatch increases. Therefore one often explicitly compensates for unseen variabilities in the acoustic space [4].

In particular, in this presentation we will introduce two techniques for unsupervised DNN speaker and environment adaptation. The first technique, termed as learning hidden unit contributions (LHUC) operates in model-space and performs adaptation by learning new combination coefficients for a speaker-independent (SI) basis (hidden units) in a speaker-dependent (SD) manner. Because the SI basis may not be optimal for unseen data, we propose a speaker adaptive trained LHUC, termed SAT-LHUC, which retains the information necessary to model the individual characteristics of the speakers in training data (not just their average aspect) and thus offers as a result more tunable canonical representation. (SAT-)LHUC operates at the level of a single hidden unit and does not allow the units to be additionally recombined with each other in a SD manner for which reason we proposed to carry the adaptation with parametric and differentiable pooling operators. More specifically, we have investigated two such parameterisations based on L_p-norm (Diff-L_p) and Gaussian (Diff-Gauss) kernels inserted in each hidden DNN layer. Parameters of such DNNs are trained in standard way using error-backpropagation and pooling parameters are later altered in a SD manner in the adaptation stage. We evaluated (SAT-)LHUC, Diff-L_p and Diff-Gauss techniques using three benchmark corpora allowing to simulate different aspects of adaptation, in particular, the amount of adaptation data per speaker, the impact of quality of both data and the associated adaptation targets, complementarity to other adaptation techniques and (for LHUC) adapting DNN models trained in sequence discriminative manner and factorisation of acoustic environments. We found the proposed techniques to improve ASR performance. On average, after LHUC and SAT-LHUC adaptation to 200 speakers of TED, AMI and Switchboard data relative WER reductions of 7.0% and 9.7% were observed with respect to SI models. Differentiable pooling versions were found to work better in the SI case and the gains from adaptation were comparable in relative terms to the ones obtained with LHUC, which was also to be found complementary in the joint (LHUC + Diff) setting.

3. References

[1] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, 2016.

[2] P. Swietojanski and S. Renals, "Differentiable Pooling for Unsupervised Acoustic Model Adaptation," *To appear in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. -, no. -, pp. 1–1, 2016.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[4] D. Yu, M. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition," in *Proc. ICLR*, 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=189337

A perceptually motivated loss function for DNN-based binary mask estimation for speech separation

Danny Websdale, Ben Milner

University of East Anglia, United Kingdom
d.websdale@uea.ac.uk, b.milner@uea.ac.uk

1. Abstract

This work proposes a perceptually motivated loss function for deep neural network (DNN) binary mask estimation for speech separation. Previous loss functions have focussed on maximising the accuracy of mask estimation but we now propose a loss function that focuses on maximising the hit minus false-alarm (HIT-FA) rate which is known to correlate more closely to speech intelligibility. Within HIT-FA a HIT refers to the proportion of correctly labeled target-dominant time-frequency units while FA refers to the proportion of incorrectly labeled noise-dominant time-frequency units. Therefore we present a loss function that maximises HITs and minimises FAs instead of the overall accuracy of the predicted masks. Evaluations on the perceptually motivated loss function show improvements to the HIT-FA rate across babble and factory noises at all signal-to-noise ratios tested.

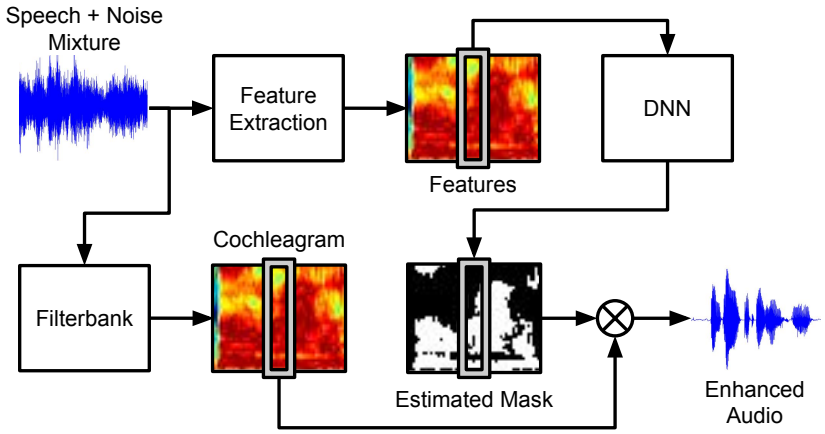


Figure 1: Overview of the speech separation system.

Hypothesis posterior student-teacher training

Jeremy H. M. Wong and Mark J. F. Gales

University of Cambridge, United Kingdom

jhmw2@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

The performance of Automatic Speech Recognition (ASR) can often be significantly improved by combining multiple systems together, especially when the quantity of training data is limited. Though beneficial, ensemble methods can be computationally expensive, often requiring multiple decoding runs. An alternative approach, appropriate for deep learning schemes, is to adopt student-teacher training. Here, a student model is trained to reproduce the outputs of a teacher model, or ensemble of teachers. The standard approach in ASR is to train the student model to emulate the frame posteriors of the teacher ensemble. This can be achieved by minimising the KL-divergence between the Deep Neural Network (DNN) outputs of the student and teacher ensemble. It is then hoped that having similar frame posteriors will result in a similar word error rate at test time. This work extends upon the current student-teacher framework, by examining the interaction between student-teacher training schemes and sequence discriminative training criteria. This may prove beneficial, as training with sequence discriminative criteria has been shown to yield significant performance gains over frame-level criteria.

This work first investigates the nature of the teacher ensemble. Diversity is introduced into the ensemble by using a different random initialisation of the DNN acoustic model for each teacher. Additionally, the teacher models can be trained with sequence discriminative criteria to further improve the ensemble recognition performance. The experiments then assess whether through frame-level student-teacher training, the student model is able to gain from these sequence-trained teachers. However, since the student model has only been trained at the frame level, training it further with sequence discriminative criteria may improve the performance.

Current student-teacher training methods only propagate information about the frame posteriors from the teacher ensemble to the student. This work expands upon this by instead propagating information about the hypothesis posteriors. As with frame-level student-teacher methods, the student model here can analogously be trained by minimising the KL-divergence between the hypothesis posteriors of the student and teacher ensemble. Training as such should allow sequential information to be taken into account, allowing student-teacher training to be directly integrated with sequence discriminative training. This proposed criterion reduces to the Maximum Mutual Information (MMI) criterion when the teacher hypothesis posterior distribution is a delta function at the reference transcription.

These approaches are evaluated on two speech recognition tasks: a Wall Street Journal (WSJ) based task and a Tok Pisin conversational telephone speech task from the IARPA Babel programme. The very limited language pack is used for Tok Pisin, containing approximately 3 hours of training data. This along with WSJ are both fairly low-resource tasks, for which ensemble methods are expected to be particularly helpful in. The results demonstrate that the gains from sequence discriminative training of the teacher ensemble can be emulated by the student model through frame-level student-teacher training. Further sequence discriminative training of the student model can bring additional gains. Finally, training the student-model using the proposed hypothesis-level criterion performs better than frame-level student-teacher training, even with subsequent sequence discriminative training of the student-model.

Poster Session 3
Tuesday June 21st, 11:45, Diamond - Workroom 1

- M. Al Dabel, J. Barker: "Time–varying Spectral Shape Optimisation Approaches for Near–end Intelligibility Enhancement"
- R. Alghady, Y. Gotoh, S. Maddock: "Analysis of visemes in the GRID corpus"
- L. Bai, P. Weber, P. Jancovic, M. Russell: "Exploring Relationships between Low–Dimensional Bottleneck Features Extracted from Neural Networks with Different Initialisations"
- P. Baljekar, A. Black: "Utterance Selection Techniques for TTS Systems Using Found Speech"
- F. Barrientos: "How categorical is categorical perception in L2? Discrimination and identification along 1–to–1 and 2–to–1 mappings"
- B. Chen, J. Lai, R. Sun, K. Yu: "Duration/Prosody Model in LSTM based TTS"
- S. R. Gangireddy, P. Swietojanski, P. Bell, S. Renals: "Unsupervised Adaptation of Recurrent Neural Network Language Models"
- P. Green, R. Marxer, S. Cunningham, H. Christensen, J. Farwer, J. Atria, F. Rudzicz, M. Yancheva, M. Malavasi, L. Desideri, A. Coy: "Remote Speech Technology for Speech Professionals – the CloudCAST Initiative"
- Y. Guo, X. Wang, C. Wu, Q. Fu, N. Ma, G. J. Brown: "A Dual–microphone based algorithm for speech source localization in reverberant environments"
- T. Hain, J. Christian, O. Saz, S. Deena, M. Hasan, R. W. M. Ng, R. Milner, M. Doulaty, Y. Liu: "webASR 2 – Improved cloud based speech technology"
- Q. Hu, J. Yamagishi, K. Richmond, K. Subramanian, Y. Stylianou: "Initial investigation of speech synthesis based on complex–valued neural networks"
- T. Ijitona, G. Di Caterina, H. Yue, J. Soragham: "Prosodic Feature Extraction for Assessment and Treatment of Dysarthria"
- K. Kyriakopoulos, K. M. Knill, M. J. F. Gales: "Automatic assessment and error detection of non–native English speech using phoneme distance features"
- T. Le Cornu, B. Milner: "Visual–to–audio: intelligible audio speech reconstruction from visual speech"
- N. Ma, G. J. Brown: "Speech localisation in a multitalker mixture by humans and machines"
- F. Schaeffler, J. Beck, S. Jannetts: "Longitudinal variation of voices – a smartphone–based field study"
- C. Seivwright, M. Russell, S. Houghton: "A comparative analysis of continuous and discontinuous implementation of a piecewise linear continuous state HMM"
- E. Vanmassenhove, J. P. Cabral, F. Haider: "Synthesis of Speech with Emotions using Sentiment Analysis"
- O. Watts, G. E. Henter, T. Merritt, Z. Wu, S. King: "From HMMs To DNNs: Where fo the Improvements Come From?"
- X. Wei, M. Russell, P. Jancovic: "Automatic Analysis of Motivational Interviewing with Diabetes Patients"
- J. Yang, A. Ragni, M. J. F. Gales, K. M. Knill: "Log–linear System Combination Using Structured Support Vector Machines"

Time-varying Spectral Shape Optimisation Approaches for Near-end Intelligibility Enhancement

Maryam Al Dabel, Jon Barker

University of Sheffield, United Kingdom

mmaldabel3@sheffield.ac.uk, j.p.barker@sheffield.ac.uk

1. Abstract

The need for improving the intelligibility of broadcast speech is being met by a recent new direction in speech enhancement: near-end intelligibility enhancement. In contrast to the conventional speech enhancement approach that processes the corrupted speech *at the receiver-side* of the communication chain, the near-end intelligibility enhancement approach pre-processes the clean speech *at the transmitter-side*, i.e. before it is played into the noisy environment.

In this work, we describe an optimisation-based approach to near-end intelligibility enhancement using an analysis-modification-synthesis framework under an energy preservation constraint. Inspired by the empirically-observed characteristics of natural speech produced in noise (known as the Lombard effect, [1]), we propose a time-varying spectral shaping method that works by defining a weight for each spectro-temporal element of the spectrum through modifying the mel-cepstral coefficients on segment-by-segment basis. Segment boundaries are selected by locating minima in the time-varying energy of the speech signal. Weights are set by optimising a measure of objective intelligibility while using a priori knowledge of the speech and noise signals.

We experiment with two contrasting measures of objective intelligibility. First, as a baseline, we implement an audibility measure named Glimpse Proportion (GP) [2]. It is defined as the percentage of spectro-temporal elements that have a local SNR higher than a given threshold. Second, we employ a discriminative microscopic intelligibility (DIS) measure [3] that is derived from a statistical model of speech from the speaker that is to be enhanced. Specifically, we employ a hidden Markov model and consider the ratio of the likelihoods of the correct state and the best scoring competing state using missing feature theory to account for masking. Whereas the GP system works by minimising the energetic masking, the DIS operates by minimising the confusions between acoustically similar speech units.

In a recent large-scale evaluation of speech near-end intelligibility enhancement techniques, Cooke et al. [4, 5] found that techniques that used a dynamic range compression (DRC) algorithm [6] showed large improvements in the considered stationary and non-stationary noises. We thus chose the full system proposed by Zorila et al. [6] as our reference system (spectral shaping and DRC). Further, our proposed approaches were combined with the time-scale modification method, i.e. DRC algorithm, for a fair comparison to the reference system.

Subjective evaluation of the proposed systems was performed for speech mixed with speech-shaped and babble-modulated noises at signal to noise ratios of -9, -6 and -3 dB. The results showed significant and consistent improvements in subjective intelligibility for the modified speech compared to the original speech. Intelligibility of DIS-based modified speech was greater than that of speech enhanced using the reference SS-DRC system in the babble-modulated noise, and was as intelligible as the reference system in the speech-shaped noise.

2. References

[1] J. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, p. 510, 1993.

[2] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[3] M. Al Dabel and J. Barker, "Speech pre-enhancement using a discriminative microscopic intelligibility model," in *Proc. Interspeech*, Singapore, 2014.

[4] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[5] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, Lyon, France, 2013, pp. 3552–3556.

[6] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, Portland, USA, 2012.

Analysis of visemes in the GRID corpus

Rabab Algadhy, Yoshihiko Gotoh, and Steve Maddock

University of Sheffield, United Kingdom

rralkasah1@sheffield.ac.uk, y.gotoh@sheffield.ac.uk, s.maddock@sheffield.ac.uk

1. Abstract

Humans have the ability to spot any slight inaccuracy in visual speech animation. Thus, creating natural-looking mouth animation remains a major challenge for developers aiming to animate a 3D talking head. The aim of our work is to use 2D audiovisual data and additional knowledge of mouth structure and movement to animate a 3D mouth and its internal features. This will involve tracking 2D features, particularly the lip shape and any visible internal mouth structures, analysing the audio signal, and mapping this data to a 3D model which makes use of visemes (where a viseme is the position of the lips, jaw and tongue when producing a particular sound). Our initial work has focussed on analysing the phonemes in the GRID audiovisual sentence corpus [1].

The GRID audio-visual dataset is provided by the University of Sheffield to support joint computational-behavioural studies in speech perception and automatic speech recognition. This corpus contains a collection of audio and video recordings of 1000 sentences produced by 34 speakers (16 female and 18 male). Each sentence contains a 6 word sequence and is formed as follows: <command: bin, lay, place, set> <color: blue, green, red, white> <preposition: at, by, in, with> <letter: A-Z excluding W> <digit: 1-9, zero> <adverb: again, now, please, soon>. Audio files have a maximum amplitude value of 1 and are downsampled to 25 kHz, with the raw original of 50kHz also being provided. Video files are sampled at 25 fps (frames per second) with normal quality formats and a frame size of 360 x 288 pixels, and high quality with a frame size of 720 x 576 pixels. The speakers face is illuminated uniformly and the background is plain blue.

Videos of one speaker delivering sentences from the GRID dataset have been used to map 42 phonemes into 16 visemes. The phoneme segmentation files of the GRID corpus that contain the sample rate of phonemes per second have been processed to determine the duration of each phoneme per frame corresponding to the video frames. For each viseme, we looked at the corresponding phonemes in the video frames to select the most closely-matching frame to be processed and used in our model (Figure 1). A table of the processed dataset has been prepared. It contains the viseme of the 3D head, the corresponding phonemes set, the number of occurrences of each phoneme set in the processed videos, and the most closely-matching frame to the 3D viseme, with its surrounding phonemes and associated video filename. In producing this dataset the following observations were made concerning coarticulation effects. The alveolar and velar visemes are harder to distinguish, because while the speaker is pronouncing them, the lips take the shape of the following viseme. This is not the case with the bilabial, Palato-alveolar and labio dental visemes, which are clear. For the vowel visemes, the rounding lips visemes are clear compared with the open lips visemes where their open mouth shapes are nearly similar.

(a) Ch,Jh,Sh viseme (b) B,M,P viseme (c) Oh viseme

Figure 1: Examples of some visemes of the GRID corpus, corresponding phonemes of viseme (a) are (\ tʃ\,\ ʃ\,\dʒ\,\ʒ\ ), (b) are (\b\,\m\,\p\) and (c) are (\ɔː\,\ɔɪ\,\əʊ\,\ aʊ\) .

2. References

[1] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

Exploring Relationships between Low-Dimensional Bottleneck Features Extracted from Neural Networks with Different Initialisations

Linxue Bai, Philip Weber, Peter Jančovič, Martin Russell

School of Engineering, The University of Birmingham, Birmingham B15 2TT, UK

lxb190@bham.ac.uk,dr.philip.weber@ieee.org,{p.jancovic, m.j.russell}@bham.ac.uk

1. Abstract

Understanding how (deep) neural networks work is an important goal. Research [1] has suggested that neurons of hidden layers in phone-discrimination neural networks are selective to phonetic features. Very low-dimensional bottleneck features (BNFs), obtained from bottleneck neural networks trained with phone posterior probability targets, have been shown to give very good speech recognition performance and preserve well the trajectory continuity for dynamic models [2]. 9 dimensional (9-dim) bottleneck features has been shown to perform similar to 39-dim conventional Mel frequency cepstral coefficients (MFCCs) in terms of GMM-HMM recognition results.

As neural networks are trained with random weight initialisations, it is interesting to ask how different initialisations affect bottleneck features. For example, do they produce similar BNFs, do the BNFs give similar recognition rate, what relationships are between these different sets of BNFs, etc.? Most neural networks are of huge size and it is difficult to investigate the differences or similarities between networks. However, our BNF are very low-dimensional features and it is comparatively easy to explore relationships between BNFs from different initialisations of the same neural network.

We analyse the similarities and differences between 9-dimensional BNFs produced from neural networks of the same structure with different random initialisations. Neural networks are trained on the TIMIT using recordings of 462 speakers, excluding the SA recordings. The input to the neural network is 26-dim logarithm filter-bank energies (logFBEs) of the current signal frame and five preceding and five following frames. The neural networks are trained with posterior probabilities of the 49 phones as targets. The neural network has been trained several times with different initial network parameters to extract multiple sets of BNFs. It is shown that the resulting sets of BNFs are different, but that they give similar phone recognition performance.

We explore the relationships between these sets of BNFs. We show that the relationship between them is approximately piecewise linear by evaluating the BNFs in a standard GMM-HMM recogniser as before. We present the results of experiments in which hierarchical clustering is applied to phone-dependent linear transformations between BNF sets. We show that the combined linear transforms that emerge correspond, in general, to phonetically meaningful phone classes. In addition, we show that the biggest decreases in phone recognition accuracy occur when transforms corresponding to categories that differ significantly in their phonetic properties are combined. This result suggests that the network is able to learn and combine multiple phone category dependent feature extraction mappings to optimise a low dimensional representation for its phone classification task.

2. References

[1] T. Nagamine, M. L. Seltzer, and M. N, "Exploring how deep neural networks form phonemic categories," in *Interspeech*, 2015, pp. 1912–1916.

[2] L. Bai, P. Jančovič, M. Russell, and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," in *Interspeech*, 2015, pp. 583–587.

Utterance Selection Techniques for TTS Systems Using Found Speech

Pallavi Baljekar and Alan Black.

Carnegie Mellon University
{pbaljeka,awb}@cs.cmu.edu.com

1. Abstract

The goal in this paper is to investigate data selection techniques for found speech. Found speech unlike clean, phonetically-balanced datasets recorded specifically for synthesis contain a lot of noise which might not get labeled well and it might also contain utterances with varying channel conditions. These channel variations and other noise distortions might sometimes be useful in terms of adding diverse data to our training set, however in other cases it might be detrimental to the system. The approach outlined in this work investigates various metrics to detect noisy data which degrade the performance of the system on a held-out test set. We assume a seed set of 100 utterances to which we then incrementally add in a fixed set of utterances and find which metrics can capture the misaligned and noisy data. We report results on three datasets, an artificially degraded set of clean speech, a single speaker database of found speech and a multi - speaker database of found speech. All of our experiments are carried out on male speakers of American English. We also show that comparable results are obtained on a female multi-speaker corpus of American English.

Found data is any data that is available readily in the public domain. This includes data from audiobooks, public speeches, news and radio broadcasts, Youtube data and telephone conversations. This type of data can either be single speaker as in the case of audiobooks and public speeches or multi-speaker as in the case of telephone conversations and news broadcasts. Thus, to build good systems from such data we would like to choose utterances which are both *representative* while also being *informative*.

In this paper we broadly categorize the error types found in such data, into two main types, *misalignment errors* and errors due to *variation in channel conditions*. Misalignment errors occur because certain types of sounds are not described in the transcript such as claps, laughs, coughs, breaths, etc. The second type of error is caused due to varying microphone conditions, channel noise etc. Errors of type 1 are never good for the system and we would like to detect and remove utterances containing such errors from our training data. However, errors of type 2 caused due to channel and speaker variation and noise, might in some cases be good for training our models, adding to the diversity. Thus our goal here is to find good measures which detect for the misalignment errors and bad utterances which will be detrimental to the system, while retaining the sentences which even though they might not be representative of the training set, might still provide valuable and diverse characteristics to the training data.

Our results indicate that it is much easier to detect misaligned data than it is to detect noisy channel variations. We also find that the data selection does scale even when 50% of the dataset has been misaligned and yields a Mean Cepstral Distortion (MCD) which is 0.3 lower than the baseline.

To select utterances we use various metrics to compare the synthesised wavefiles with respect to the original wavefiles. All of our metrics are calculated at the utterance level. The spectral measures include mean cepstral distortion, instantaneous frequency, global variance and modulation spectrum based measures as well as various cross correlation based measures calculated by applying the Teager Energy operator to the waveform. We also look at differences in duration prediction as a criteria for utterance selection.

From all of the measures we explored, we find that the mean cepstral distortion performs the best followed by the error in the duration prediction. We surprisingly find that the various cross-correlation based metrics are not good indicators of the presence of misaligned data. In addition, contrary to our expectations, we find the global spectral measures, i.e., modulation spectrum and global variance perform poorly in detecting changes in channel conditions.

In this paper, we show that our claim that not all data is good data holds. We show that selecting a smaller, cleaner subset for voice building is much better and less time consuming than building from the full noisy dataset. We explore various utterance level metrics which would be indicators of the *measure of goodness* of an utterance. We show results considering both the availability of a small seed set of about 100 utterances and building from all of the noisy data without access to such a seed dataset. We also explore and contrast the results of iteratively realigning the data each time as opposed to using all of the data for alignment and only using metrics to select utterances based on this initial alignment.

We show preliminary results in this paper on data selection. We show that it scales well on misaligned data and gives us significantly better results than using the entire subset of noisy data. In the future, we would like to test our methods on large non-English corpora such as the Babel datasets, that contain speech recorded over telephones, with the main aim of building understandable speech synthesis systems for low resource languages.

How categorical is categorical perception in L2? Discrimination and identification along 1-to-1 and 2-to-1 mappings

Fernanda Barrientos

University of Manchester, United Kingdom

fernanda.barrientoscontreras@manchester.ac.uk

1. Abstract

Perception of nonnative contrasts in L2 has been a widely productive topic in second language phonology, but the extent to which these newly created perceptual categories differ from L1 perceptual categories remains unclear. From an experimental perspective, the presence of phonemic representations allow for speakers to show categorical perception, a phenomenon consisting of a correspondence between labelling and discrimination of tokens along a synthesized continuum between two different sounds with a phonemic status in the speakers grammar [1]. Although this categorical perception pattern does not seem to be as consistent for vowels than for consonants [2], it seems to be the only experimental approach that shows the presence of phonemic categories from a perceptual point of view. Furthermore, this approach has not been used in L2 phonology, given the bias of the labelling task.

This work aims to observe the relation between labelling and discrimination of two L2 vowels that are mapped onto the same L1 category, and draw conclusions regarding the status of these newly created perceptual categories in the L2 speakers phonological knowledge. The experiment is an extension of the Liberman et al. approach and it consists of a discrimination task and two different types of labelling: one with L2-like labels and another one with L1-like labels. Subjects were 8 native speakers of Spanish with high English proficiency, 8 native speakers of Spanish with low proficiency, and a control group of 8 native speakers of English, who were asked to label and discriminate tokens of vowels along the /ɑ-ʌ/ and the /ʌ-ɛ/ continua. While the L2-like labels (is this a vowel like the one in the word cup?) aim to test the presence of a sound category that does not rely on any type of L1 transfer, the L1-like language mode and labels (es esta una vocal como la de la palabra mar?) are looking for an explanation to possible miscategorization and/or reduced discrimination.

Preliminary results for the ɑ-ʌ continuum show that native speakers were able to achieve perfect categorical labelling pattern, but below chance 1-step discrimination and a significant increase in accuracy at the boundary zone in 2-step discrimination. Beginners showed random discrimination and labelling, and advanced speakers presented a mismatch between labelling and discrimination, where L2 labels are assigned at above chance level, but discrimination is below chance. L1-like labelling results, on the other hand, show that tokens of both /ɑ/ and /ʌ/ are perceived as the Spanish vowel /a/. Conversely, results for the /ʌ-ɛ/ continuum were very similar (categorical labelling and sensitivity peak in discrimination at the boundary) across the three groups, with L1 labels showing that /ʌ/ and /ɛ/ are mapped onto different L1 categories: Spanish /a/ and /e/. These results suggest that even fully proficient L2 speakers do not store representations of sounds in the same way as L1 sounds; rather, when L2 categories are mapped onto the same perceptual L1 category, labelling takes place in a probabilistic manner and sensitivity peaks in discrimination are unavailable due to the absence of phonemic categories that warp the perceptual space [3].

2. References

[1] A. Liberman, K. Harris, H. Hoffman, and B. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, no. 5, pp. 358–368, 1957.

[2] E. Gerrits and M. Schouten, "Categorical perception depends on the discrimination task," *Perception and Psychophysics*, vol. 66, no. 3, pp. 363–376, 2004.

[3] P. Kuhl and P. Iverson, "Chapter 4: Linguistic experience and the "perceptual magnet effect"," *Speech perception and linguistic experience: Issues in cross-language research*, pp. 121–154, 1995.

Duration/Prosody Model in LSTM based TTS

Bo Chen, Jiahao Lai, Ruihua Sun, Kai Yu

Shanghai Jiao Tong University, China

bobmilk@sjtu.edu.cn

1. Abstract

Prosody, commonly consisting of `Duration` and `F0`, is one of the major facts that affect the speech naturalness. Different from `F0`, `Duration` is always modelled independently from acoustic features in Neural Network based statistical parameter speech synthesis framework. This ongoing work is to investigate how should we make use of duration infomation in LSTM acoustic modelling and how the duration accuracy affects the performance in LSTM acoustic model with duration information at different level. Serveral LSTM based duration models are compared. The generated speech from a very simple LSTM acoustic model are evaluated in listening test. This work will be submitted to SLT2016.

Table 1: *Objective Evaluation of Acoustic Models with Duration Information at different level.*

Corpus	Dur in AM	F0 RMSE	VCE(%)	MCD
xijunm	phone	12.00	4.33	5.97
	state	12.02	3.52	5.69
	rich	**11.69**	**3.23**	**5.52**

Table 2: *Root Mean Square Error of duration models.*

Corpus	Level	Tasks	dur(s)	dur(p)	log dur(p)
xijunm chn 13h	gmm	—	3.29	4.98	0.294
	phone	dur(p)	—	4.77	0.276
	phone	dur(sp)	3.21	4.65	0.271
	state	dur(s)	3.20	4.70	0.277
	state	dur(sp)	3.17	4.42	0.257
	state	dur(sp),lf0	**3.16**	**4.38**	**0.256**

SJTU SPEECH LAB

上海交通大學智能語音實驗室

Unsupervised Adaptation of Recurrent Neural Network Language Models

Siva Reddy Gangireddy, Pawel Swietojanski, Peter Bell and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

s.gangireddy@sms.ed.ac.uk

1. Abstract

Recurrent neural network language models (RNNLMs) have been shown to consistently improve Word Error Rates (WERs) of large vocabulary speech recognition systems employing n-gram LMs. In this paper we investigate supervised and unsupervised discriminative adaptation of RNNLMs in a broadcast transcription task to target domains defined by either genre or show. We have explored two approaches based on (1) scaling forward-propagated hidden activations (Learning Hidden Unit Contributions (LHUC) technique [1]) and (2) direct fine-tuning of the parameters of the whole RNNLM. To investigate the effectiveness of the proposed methods we carry out experiments on multi-genre broadcast (MGB) data following the MGB-2015 challenge protocol. We observe small but significant improvements in WER compared to a strong unadapted RNNLM model.

2. References

[1] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. IEEE Workshop on Spoken Language Technology*, Lake Tahoe, USA, Dec. 2014.

Remote Speech Technology for Speech Professionals – the CloudCAST Initiative

Phil Green[1], Ricard Marxer[1], Stuart Cunningham[1], Heidi Christensen[1], Jochen Farwer[1], Jose Atria[1]
Frank Rudzicz[2], Maria Yancheva[2], Massimiliano Malavasi[3], Lorenzo Desideri[3], Andre Coy[4]

[1]CATCH, University of Sheffield, UK
[2]University of Toronto and Toronto Rehabilitation Institute, Canada
[3]AIAS Onlus Bologna, Italy
[4]University of the West Indies, Jamaica

1. Abstract

Clinical applications of speech technology face two challenges. The first is data sparsity. There is little data available to underpin techniques which are based on machine learning and, because it is difficult to collect disordered speech corpora, the only way to address this problem is by pooling what is produced from systems which are already in use. The second is personalisation. This field demands individual solutions, technology which adapts to its user rather than demanding that the user adapt to it. Here we introduce a project, CloudCAST, which addresses these two problems by making remote, adaptive technology available to professionals who work with speech: therapists, educators and clinicians.

A Dual-microphone based algorithm for speech source localization in reverberant environments

Yanmeng Guo[1]*, Xiaofei Wang*[1]*, Chao Wu*[1]*, Qiang Fu*[1]*, Ning Ma*[2]*, Guy J. Brown*[2]

[1]Institute of Acoustics, Chinese Academy of Sciences
[2]Department of Computer Science, University of Sheffield
guoyanmeng@mail.ioa.ac.cn, {wangxiaofei, wuchao, qfu}@hccl.ioa.ac.cn
{n.ma, g.j.brown}@sheffield.ac.uk

1. Abstract

Speech source localization (SSL) using a microphone array aims to estimate the direction-of-arrival (DOA) of the speech source. However, its performance often degrades rapidly in reverberant environments. A novel dual-microphone SSL algorithm is proposed to address this problem. First, the time-frequency regions dominated by direct sound are extracted by tracking the envelope of speech, reverberation and background noise. The time-difference-of-arrival (TDOA) is then estimated by considering only these reliable regions. Second, a bin-wise de-aliasing strategy is introduced to make better use of the DOA information carried at high frequencies, where the spatial resolution is higher and there is typically less corruption by diffuse noise. Compared with other widely-used algorithms, the proposed algorithm shows more reliable performance in realistic reverberant environments.

SSL is important for voice capture in many human-computer interaction applications, such as human-robot interaction, camera steering and intelligent monitoring. Generally, the far-field assumption is applicable for a small scale microphone array, so that the DOA can be estimated from the time difference of arrival (TDOA) or synchrony between the received signals. Most SSL algorithms are reliable in free-field conditions, in which the received signal contains only the direct wave of the speech. However, in real application environments, where room reflections occur, their performance degrade inevitably because the captured signal contains both the direct sound and reverberation.

To achieve robustness in the presence of reverberation, two improvements are proposed here to extract the reliable parts of the received signal. First, the received signal is looked as the summation of three components: direct speech, reverberation and ground noise, and the time-frequency regions dominated by direct speech are extracted through an envelope tracking strategy, in which the characteristics of the speech signal, reverberation and background noise are exploited. Second, a bin-wise de-aliasing process is proposed to make better use of the TDOA information carried in high frequency parts. Generally, the high frequency parts are less corrupted by reverberation, because on average high frequency signal has a higher absorption ratio. This process is based on the assumption that only one active speech source exists.

The performance of the proposed algorithm is compared with four widely-used algorithms: Generalized Cross Correlation (GCC), GCC with phase transform (GCC-PHAT), Steered Response Power (SRP) and SRP with phase transform (SPR-PHAT). The test corpus is recorded in a $6\,\mathrm{m} \times 5\,\mathrm{m} \times 3\,\mathrm{m}$ varechoic chamber with T_{60} varies from 300 ms to 700 ms. All the algorithms have low bias in the DOA$= 0^\circ$ condition, regardless of whether the reverberation is high or low. However, the performance degrades when DOA or reverberation becomes higher. The proposed algorithm shows the lowest bias when the DOA is not 0°, and the bias increases much more slowly with DOA and reverberation level.

There are still some limitations of this algorithm. First, the de-aliasing process is based on the assumption that there exists only one nonstationary sound source, and this condition is not always applicable in real applications. Second, the envelope tracking process is deployed in each frequency bin, and the correlation between different frequency bins could be further exploited. In both respects, a strategy that groups correlated frequency bins will be helpful. Instead of the separate envelope tracking in each frequency bin, a contour tracking that involves several correlated frequency bins will be more practical in multi-source conditions to extract the direct sound, thus allowing the spatial de-aliasing strategy to be generalized to deal with multi-source conditions. These will be addressed in our future research.

webASR 2 - Improved cloud based speech technology

Thomas Hain, Jeremy Christian, Oscar Saz, Salil Deena, Madina Hasan, Raymond W. M. Ng, Rosanna Milner, Mortaza Doulaty, Yulan Liu

Speech and Hearing Research Group, The University of Sheffield, UK

{t.hain, jchristian1, o.saztorralba, s.deena, m.hasan, wm.ng, rmmilner2, mortaza.doulaty, yulan.liu}@sheffield.ac.uk

1. Abstract

This paper presents the most recent developments of the webASR service (www.webasr.org), the world's first web–based fully functioning automatic speech recognition platform for scientific use. Initially released in 2008, the functionalities of webASR have recently been expanded with 3 main goals in mind: Facilitate access through a RESTful architecture, that allows for easy use through either the web interface or an API; allow the use of input metadata when available by the user to improve system performance; and increase the coverage of available systems beyond speech recognition. Several new systems for transcription, diarisation, lightly supervised alignment and translation are currently available through webASR. The results in a series of well–known benchmarks (RT'07, RT'09, IWSLT'12 and MGB'15 evaluations) show how these webASR systems provides state–of–the–art performances across these tasks.

Table 1: *Benchmark results for webASR systems.*

Transcription systems					
System	Benchmark	Substitutions	Deletions	Insertions	WER
Meeting	RT'09	18.4%	6.8%	3.3%	**28.5%**
Lectures	IWSLT'12	8.0%	2.3%	2.6%	**12.9%**
Media	MGB'15	14.1%	10.7%	3.2%	**28.0%**
Segmentation and diarisation systems					
System	Benchmark	Missed speech	False alarm	Speaker error	SER/DER
Meeting	RT'07	11.8%	10.7%	-	**22.5%**
Media	MGB'15	1.9%	6.4%	41.1%	**49.3%**
Lightly supervised alignment system					
System	Benchmark		Precision	Recall	F–measure
Media	MGB'15		0.8818	0.8689	**0.8753**
Machine translation system (English to French)					
System	Benchmark			WER(English)	BLEU(French)
Lectures	IWSLT'12			12.5%	**31.28**

Initial investigation of speech synthesis based on complex-valued neural networks

Qiong Hu[1], *Junichi Yamagishi*[1], *Korin Richmond*[1], *Kartick Subramanian*[2], *Yannis Stylianou*[3]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]Toshiba Research Europe Ltd, Cambridge, U.K.

1. Abstract

For many real-valued signals, one of the most frequently used approaches is frequency domain analysis such as the Fourier transform, which normally leads us to a complex domain. The statistical behaviour and properties of amplitude spectra and related parameterizations are well known and have been used in many speech processing applications. Various models have been proposed to model the statistical behaviour of these parameters [5]. Meanwhile, recent studies [3] have elaborated the potential of using phase features and the common strategy among these methods is to analyse and model the amplitude and phase separately [2].

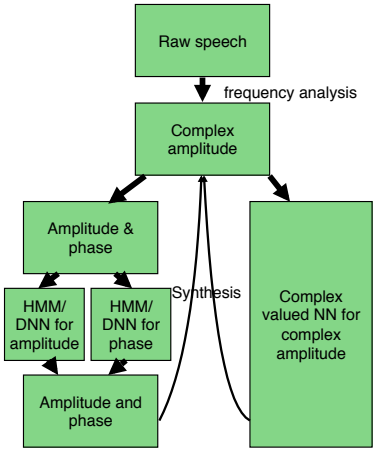


Figure 1: Comparison of traditional (left) and proposed systems (right) for amplitude and phase modelling

An alternative approach to such explicit and separated amplitude and phase feature representations is to combine amplitude and phase together by representing a signal as a complex value $z = u + iv \in C$, and then to model the signal z using a new statistical model, which can deal with complex numbers directly. Here we may use both the amplitude and phase information of the signal as a part of the new objective function in the complex domain $E_C(z) = \hat{E}_C(A, \varphi)$ for learning the models so that the model can consider errors of the amplitude A and phase φ of the signal z jointly (Figure 1).

In this poster, an alternative model referred to as complex valued neural network (CVNN) [1, 4] is applied for SPSS. A complex exponential function, which has singularity points at $\pm\infty$ only is used at the output layer while the $Sinh$ is used as hidden activation function. A complex-valued back-propagation algorithm using a logarithmic minimization criterion which includes both amplitude and phase errors is used as a learning rule. In this preliminary work, three parameterization methods are studied for mapping text to acoustic features: cepstrum / real-valued log amplitude, complex amplitude with minimum phase and complex amplitude with mixed phase. Our results show the potential of using CVNN for modelling both real and complex valued acoustic features.

2. References

[1] I. Aizenberg. *Complex-valued neural networks with multi-valued neurons*, volume 353. Springer, 2011.

[2] R. Maia and Y. Stylianou. Complex cepstrum factorization for statistical parametric synthesis. In *Proc. ICASSP*, 2014.

[3] P. Mowlaee, R. Saeidi, and Y. Stylanou. INTERSPEECH 2014 special session: Phase importance in speech processing applications. In *Proc. Interspeech*, 2014.

[4] S. Suresh, N. Sundararajan, and R. Savitha. *Supervised learning with complex-valued neural networks*. Springer, 2013.

[5] H. Zen. Acoustic modeling in statistical parametric speech synthesis–from HMM to LSTM–RNN. In *Proc. MLSLP*, 2015.

Prosodic Feature Extraction for Assessment and Treatment of Dysarthria

Tolulope Ijitona, Dr Gaetano Di Caterina, Dr Hong Yue, Professor John Soraghan

Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom
tolulope.ijitona@strath.ac.uk

1. Abstract

Dysarthria, a neurological motor speech disorder caused by lesions to the central and peripheral nervous system, accounts for over 40% of neurological disorders referred to pathologists in 2013[1]. This affects the ability of speakers to control the movement of speech production muscles due to muscle weakness. Dysarthria is characterised by reduced loudness, high pitch variability, monotonous speech, poor voice quality and reduced intelligibility [2]. Current techniques for dysarthria assessment are based on perception, which do not give objective measurements for the severity of this speech disorder. There is therefore a need to explore objective techniques for dysarthria assessment and treatment.

The goal of this research is to identify and extract the main acoustic features which can be used to describe the type and severity of this disorder. An acoustic feature extraction and classification technique is proposed in this work. The proposed method involves a pre-processing stage where audio samples are filtered to remove noise and resampled at 8 kHz. The next stage is a feature extraction stage where pitch, intensity, formants, zero-crossing rate, speech rate and cepstral coefficients are extracted from the speech samples. Classification of the extracted features is carried out using a single layer neural network. After the classification, a treatment tool is to be developed to assist patients, through tailored exercises, to improve their articulatory ability, intelligibility, intonation and voice quality.

Consequently, this proposed technique will assist speech therapists in tracking the progress of patients over time. It will also provide an acoustic objective measurement for dysarthria severity assessment. Some of the potential applications of this technology include management of cognitive speech impairments, treatment of speech difficulties in children and other advanced speech and language applications.

2. References

[1] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*: Elsevier Health Sciences, 2013.

[2] J. P. Hosom, A. B. Kain, T. Mishra, J. P. H. v. Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. I-924-I-927 vol.1.

Automatic assessment and error detection of non-native English speech using phoneme distance features

Konstantinos Kyriakopoulos, Kate M. Knill, Mark J.F. Gales

ALTA Institute / Department of Engineering
University of Cambridge, United Kingdom

kk492@cam.ac.uk, kmk1001@cam.ac.uk, mjfg@eng.cam.ac.uk

1. Abstract

With growing global demand for learning English as a second language, there has been considerable interest in methods of automatic evaluation of spoken language proficiency for use in interactive electronic learning tools as well as for grading candidates for formal qualifications. A system is presented for automatically grading the fluency level of non-native English speakers and identifying individual pronunciation errors in their utterances, using only short samples of unstructured, spontaneous speech.

The system builds on the baseline grader system [1] developed at the ALTA Institute (Figure 1) in which audio is passed through an automatic speech recogniser (ASR), and the recognised text and audio used to extract prosodic and other fluency features, which are in turn used to train a Gaussian Process (GP) grader to assign scores to speakers and sections, based on training data obtained from a corpus of recorded answers to the BULATS English speaking test [2], labeled with overall speaker and section-specific scores.

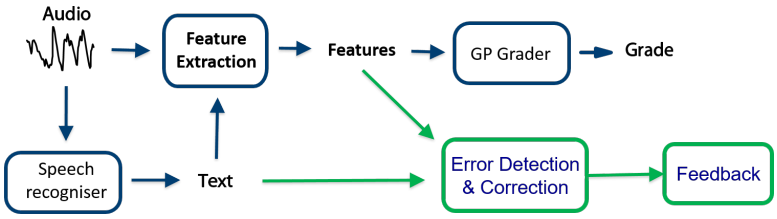


Figure 1: Outline of the ALTA grading system

The grader is enhanced by extracting new *phoneme distance features* to represent relative phoneme pronunciation within a certain context (e.g. for a certain speaker or in a certain utterance). They are computed as the K-L divergences between each possible pair of a set of models trained to represent the manner of pronunciation of each of the 47 phonemes in the English language within that context. Two different types of model are investigated to represent manner of pronunciation: a simple multivariate Gaussian with a diagonal covariance matrix and a three-emitting state Hidden Markov Model (HMM) with two-component diagonal covariance matrix Gaussian Mixture Model (GMM) states, the latter obtained using CMLLR maximum likelihood model adaptation. The features obtained are then added to the ALTA speaker and section graders, producing considerable performance enhancement for both speaker native language (L1) dependent and independent paradigms, though more so for the former (see Table 1).

L1	baseline feats.	pron. feats.	baseline + pron.
Gujarati	0.816	0.843	0.871
Mixed	0.807	0.833	0.840

Table 1: Pearson correlation between best automatic grader speaker scores and corresponding human expert scores, using baseline features only, phoneme distance pronunciation features only and all features together

The same features are then used to train individual GP scorers for each pair of phonemes, capturing the relationship between each K-L divergence and score. Models adapted to the speakers being tested can now be further adapted to the level of individual utterances and the K-L divergences between them used to predict scores for each phoneme pair, which are in turn used to derive phoneme scores by inverse variance weighting. These scores are then used along with the ASR word confidences to identify individual phoneme level errors in the candidate's speech. The same method applied at the speaker rather than utterance level allowed identification of the phonemes most commonly mispronounced by each speaker. A method is defined to detect the precision (i.e. fraction of identified errors that are correct) of these error detection mechanisms, using crowd-sourced word-level human binary judgments of pronunciation quality.

2. References

[1] R. C. van Dalen, K. M. Knill, and M. J. Gales, "Automatically grading learners english using a gaussian process," 2015, aLTA Institute / Department of Engineering, University of Cambridge.

[2] BULATS. Business language testing service. [Online]. Available: http://www.bulats.org/computer-based-tests/online-tests

Visual-to-audio: intelligible audio speech reconstruction from visual speech

Thomas Le Cornu, Ben Milner

University of East Anglia, United Kingdom

t.le-cornu@uea.ac.uk, ben.milner@uea.ac.uk

1. Abstract

The aim of this work is to produce an intelligible audio speech signal from visual speech information. The primary application is for surveillance scenarios where a video signal of a speaker is available, but no audio is present. A lack of audio may be due to the distance of the recording device to the target in question, highly corrupted audio, or simply no audio recording device being present in the system.

Previous work focused on using Gaussian mixture models and deep neural networks for regression to predict the real-valued Mel-filterbank channel amplitudes and Linear Predictive Coding coefficients [1]. The visual features tested were derived from the two-dimensional discrete cosine transform and active appearance models. Various combination of statistical models, audio features, and visual features were explored to find those that produced the most intelligible speech. The best combinations were evaluated with subjective listening tests to obtain an intelligibility score. It was found that the audio speech was not of sufficient intelligibility to significantly improve upon visual-only accuracies.

In this work, the visual-to-audio mapping is reformulated as one of clustering and then classification. Audio feature vectors are clustered using the mini-batch variant of the standard k-means algorithm to produce a codebook. Labels are assigned to input audio feature vectors based on the closest cluster (lowest Euclidean distance) in the codebook. From the joint visual feature vector and cluster label, classification using neural networks can then be performed. Mel-filterbank channel amplitudes and active-appearance model based features are used for the audio and visual feature representations respectively.

As speech is a dynamic process, two methods of incorporating longer-range temporal information are explored. Firstly, at the feature level where multiple frames of contiguous audio feature vectors are clustered and then predicted from varying window widths of grouped visual feature vectors. Standard deep neural network models are used for this configuration. Secondly, at the model level by using the long short-term memory recurrent neural network architecture to learn the temporal dynamics of the audio and visual features.

Subjective listening tests are still being conducted currently, however, preliminary results suggest that audio-only intelligibility scores are vastly improved, with an expected further increase for audiovisual configurations.

2. References

[1] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 3355–3359.

Speech localisation in a multitalker mixture by humans and machines

Ning Ma and Guy J. Brown

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

{n.ma, g.j.brown}@sheffield.ac.uk

1. Abstract

In the 1950s, Cherry [1] noted the ability of listeners to attend to one speaker in the presence of others, and called this the 'cocktail party problem'. Since then, this aspect of human hearing has been the subject of much psychophysical investigation [2], and has also motivated computational work which aims to build voice separation systems. However, developing a system which matches human performance in the cocktail party problem has proven to be very challenging.

A recent psychophysical study has shown that listeners are able to exploit prior knowledge of the masker locations in a cocktail party scenario. Kopco et al. [3] investigated the ability of listeners to localise a female target voice in the presence of four spatially distributed male masking voices. They found that listeners were better able to localise the target when the spatial locations of the masker voices were cued before the task.

In this study, this effect was investigated via speech localisation experiments with both human listeners when listening over headphones and a machine system. Two configurations were used: either the masker locations were fixed or the locations varied from trial-to-trial. Performance was examined in both anechoic and reverberant conditions. Listeners are able to exploit prior information about masker locations in Kopco et al.'s task when listening over headphones, but only in reverberant conditions and when the target speech was not co-located with a masker. The effect for headphone listening was smaller than that previously reported for listening in a real room [3], but the performance pattern was similar.

The machine system uses deep neural networks (DNNs) to learn the relationship between binaural cues and source azimuth [4], and exploits top-down knowledge about the spectral characteristics of the target source and the prior knowledge of masker positions when available [5]. The auditory front-end consisted of a bank of 32 overlapping Gammatone filters, with centre frequencies uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 8 kHz. Inner-hair-cell processing was approximated by half-wave rectification. Following this, the cross-correlation between the right and left ears was computed independently for each frequency channel. As in [4], the system used the whole cross-correlation function, instead of interaural time difference (ITD), as localisation features. This approach was motivated by observations that computation of ITD may not be robust in the presence of multiple talkers, and that there are systematic changes in the cross-correlation function with source azimuth.

The computational model was able to match human data to some extent by exploiting the sources of knowledge available to listeners in this scenario, i.e. speaker characteristics and masker locations. Our experiments also show that the machine system outperformed listeners in most listening conditions.

Future work will assess the benefit of individualised head related transfer functions (HRTFs) in this task. The role of head movements will also be investigated, thus allowing listener performance to be compared with a DNN-based localisation system that uses head movements [4].

2. References

[1] C. Cherry, "Some experiments on the recognition of speech with one and two ears," *J. Acoust. Soc. Am.*, vol. 24, pp. 554–559, 1953.

[2] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[3] N. Kopco, V. Best, and S. Carlile, "Speech localization in a multitalker mixture," *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 1450–1457, 2010.

[4] N. Ma, G. J. Brown, and T. May, "Robust localisation of multiple speakers exploiting deep neural networks and head movements," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3302–3306.

[5] N. Ma, G. J. Brown, and J. A. Gonzalez, "Exploiting top-down source models to improve binaural localisation of multiple sources in reverberant environments," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 160–164.

Longitudinal variation of voices – a smartphone-based field study

Felix Schaeffler, Janet Beck, Stephen Jannetts

Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University Edinburgh
fschaeffler@qmu.ac.uk,jbeck@qmu.ac.uk, sjannetts@qmu.ac.uk

1. Abstract

Behavioural voice problems and disorders are a long-standing issue in many professions, for example teachers and call-centre workers [1;2]. While the problem has repeatedly been recognised, effective remedies are still missing, although there are clear indications that behavioural voice disorders are potentially preventable through appropriate and timely intervention [2].

Modern smartphones offer entirely new possibilities for data capture and user-controlled monitoring, but so far have not been used for voice health-related monitoring to a large extent. There are apps that provide access to measurements of acoustic parameters like shimmer, jitter or F0 range [3], but these mainly focus on a clinical context.

In our view there is great potential for voice monitoring as a preventative tool. However, this application requires increased knowledge about longitudinal variation patterns in healthy, disordered and 'at-risk' voices, to better understand which patterns of acoustic voice variation are typical, innocuous or even positive, and which patterns of variation could indicate potentially worrying changes. Furthermore, limits and constraints of smartphone based field recordings for voice parameter extraction need to be analysed, in order to understand – and if possible quantify – the influence of environmental factors like room size/room reverberation, background noise, microphone distance and smartphone type.

The voicecheck project (website: https://voicecheck.org.uk) aims at developing a database of longitudinal recordings of voices with smartphones, initially from speakers with healthy voices, in order to analyse longitudinal variation, identify potentially diagnostic variation patterns and develop dynamic thresholds for selected acoustic voice parameters.

The voicecheck smartphone app has been designed to collect voice recordings from a large range of smartphone systems. The app is HTML5 based, and free versions for iOS and Android are available in the respective app stores. The app prompts the user to record (1) two sustained [a] vowels at a comfortable pitch and loudness level, (2) nine sentences that focus on various potential voice diagnostics and (3) a short story.

Alongside each audio recording, the user also records self-reported voice use (7-item numbered rating scale), self-reported stress level (7-item numbered rating scale), self-reported recording room size and type (9-item word scale with room examples and approximate sizes), 4 variables for self-assessed voice quality, 5 variables for vocal tract/throat sensations and a variable to report the presence of a cold or similar illness (4-item word scale). Audio files are stored as uncompressed 44kHz pcm (wav) files and all data is securely transmitted to a central server for further processing and analysis after each recording.

Participant sign-up is completed online via the project website (voicecheck.org.uk). The signup procedure incorporates creation of a login, the completion of a consent form and background questionnaire, and gives access to recording instructions, the app password and a personalised recording event schedule, which is automatically emailed to the user as a calendar (ics) file that can be imported into many popular calendar apps. Participants use the voicecheck app to record themselves on up to 50 occasions. They are encouraged to choose environments with low background noise and are instructed to control microphone distance by holding the smartphone a handspan away from their mouth.

For acoustic analysis, the data is initially spliced into three parts: (1) sustained vowels, (2) voiced portions of the connected speech parts, and (3) silent portions of the signal. Identification of voiced and silent portions largely follows the procedure described in [4]. Our analysis currently focuses on the relation between reported voice use, room size/type, presence of a cold or similar illness and acoustic parameters like F0, AVQI and CPPS [4]. A pilot study of two participants showed significant increases of mean F0 in connected speech with increased voice use, a significant decrease of F0 with the presence of a cold and speaker-specific effects of voice use on CPPS.

At the UK Speech conference we will present results from additional participants and a wider range of acoustic parameters, and discuss data reliability, relevance for voice health monitoring and next analytical steps.

Acknowledgements

This project was part-funded by a Research Incentive Grant from the Carnegie Trust for the Universities of Scotland (Ref. Nr. 70230). We would like to thank Dr Matthias Eichner for crucial support with app development.

2. References

[1] N. Roy, R.M. Merrill, S. Thibeault, R.A. Parsa, S.D. Gray, E.M. Smith, "Prevalence of voice disorders in teachers and the general population", Journal of Speech, Language, and Hearing Research 47(2), 2004, pp. 281-93.

[2] L. Lehto , P. Alku , T. Bäckström , E. Vilkman, "Voice symptoms of call-centre customer service advisers experienced during a work-day and effects of a short vocal training course." *Logopedics phoniatrics vocology* 30.1, 2005, pp. 14-27.

[3] M. Mat Baki, G. Wood, M. Alston, P. Ratcliffe, G. Sandhu, J.S. Rubin, M.A. Birchall, "Reliability of operavox against multidimensional voice program (MDVP)". *Clinical Otolaryngology* 40(1), 2015, pp. 22-8.

[4] Y. Maryn, D. Weenink, "Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index." *Journal of Voice* 29(1), 2015, pp. 35-43.

Abstract Submission to UK Speech 2016

Chloe Seivwright, Martin Russell, Steve Houghton

University of Birmingham, United Kingdom

cas390@bham.ac.uk, M.J.Russell@bham.ac.uk, shoughton@bham.ac.uk

1. Abstract

This work will present the results of implementing a particular case of a probabilistic-trajectory segmental model characterised by a continuous state space. This model of speech is motivated by the underlying speech production process, such that when speech is produced our articulators move along constrained continuous trajectories. Conventional HMM systems tend to disregard this property of speech production consequently assuming independence between adjacent features of the data. By considering a more faithful model of speech that incorporates the intuitive continuous nature of speech, we aim to address the inconsistencies that arise when using a discrete state space to model speech.

There have been a number of acoustic models proposed to address the need for more accurate models of speech dynamics via segmental modelling, e.g. [1], [2]. Also, by using the Holmes Mattingly and Shearme speech synthesis model [3] in which the speech signal is approximated as a sequence of connected dwell-transition regions, for recognition using a continuous state HMM (CS-HMM) where continuous state refers to the continuous state space [4], [5]. Following this ethos we look at the effect of explicitly including a continuity constraint between adjacent segments. We will present initial results for a TIMIT experiment using a single state and a three state model. A comparative HTK experiment shows compatible results when compared with the Piecewise Linear CS-HMM when carefully limiting the number of parameters to match the CS-HMM.

On analysing the output precision matrices of these experiments via a statistical binomial significance test, it is possible to identify some interesting trends. The continuous piecewise linear decoder (PLC-Decoder) outperforms the piecewise linear decoder (PL-Decoder) on the majority of consonant bursts and closures, whereas the PLC-Decoder performs better on *all* of the voiced phonemes. The results of this experiment is consistent with our prior understanding, e.g. Abrupt changes in energy between consonants compared with smooth changes between vowels. However, the similarities in the accuracy encourages us to look more closely at how these systems work and poses the question of whether we can implement a system that relaxes the continuity constraint so that discontinuities can be accommodated in regions where expected.

By looking at the spectrogram of a TIMIT utterance we are further encouraged by the visual analysis and conclusions that can be drawn from the output trajectories of both the PL and PLC Decoders. The main focus of this poster will be a discussion as to how we intend to relax the continuity constraint of the system by factoring in a probabilistic distribution that characterises how discontinuous you expect a model to behave given its context. We will outline the problems faced when considering a bigram language model with respect to the increased parameters, and a proposal for how the system can approximate the bigram statistics during the decode instead of requiring the measured statistics that consequently requires more data to train the system parameters.

2. References

[1] W. J. Holmes and M. J. Russell, "Probablistic-trajectory segmental HMMs," *Comp. Speech & Lang.*, vol. 13, no. 1, pp. 3–37, 1999.

[2] P. J. B. Jackson and M. J. Russell, "Models of speech dynamics in a segmental-HMM recogniser using intermediate linear representations," in *Proc. Int. Conf. on Spoken Lang. Proc.,* Denver, CO, 2002, pp. 1253–1256.

[3] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech synthesis by rule," *Language & Speech*, vol. 7, pp. 127–143, 1964.

[4] C. J. Champion and S. M. Houghton, "Application of continuous state hidden markov models to a classical problem in speech recognition," *Comp. Speech & Lang.*, no. 0, pp. –, 2015, accepted for publication. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230815000443

[5] P. Weber, S. Houghton, C. Champion, M. Russell, and P. Jančovič, "Trajectory Analysis of Speech using Continuous State Hidden Markov Models," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, 2014, pp. 3042–3046.

Synthesis of Speech with Emotions using Sentiment Analysis

Eva Vanmassenhove[1]*, João P. Cabral*[2]*, Fasih Haider*[2]

[1]ADAPT, Dublin City University, Ireland
[2]ADAPT, Trinity College Dublin, Ireland
vanmassenhove.eva@gmail.com, cabralj@tcd.ie, haiderf@tcd.ie

1. Abstract

Emotional Text-To-Speech (TTS) is a challenging but important part in speech synthesis since rendering emotion makes speech sound more natural [1]. This work focuses on emotional TTS for storytelling of fairy tales in audiobooks. Rendering emotions in TTS is, however, not trivial. It requires solving two main problems: (a) predicting the correct emotional values of a sentence or utterance and (b) modelling and generating emotional speech [2]. In this work, we developed a method to predict the emotion from a sentence so that we can convey it through the synthetic voice. It is based on an existing sentiment analysis technique (detects positive/negative polarity), which was extended to detect the emotional value of a phrase for six basic emotions.

Many research works focus on emotional speech synthesis [3] or sentiment analysis [4]. However, research work on the application of sentiment analysis to TTS appears to be less common. In [5], sentiment analysis is used as an input feature for expressive speech synthesis but it is only used to distinguish between different sentiment polarities (positive, negative and neutral). The expressiveness in audiobooks has a rich variety [6] and we believe that a more fine-grained distinction between emotions is necessary to better model these speech variability factors. For example, emotions belonging to the same polarity, such as 'anger' and 'sadness' (negative polarity), are characterised by different acoustic properties (intensity, pitch, speech rate, etc.), which should be modelled by the TTS system. In this work, we propose a novel emotion labelling system that uses both the information of the emotional polarity from sentiment analysis [7] and emotion category from the NRC Emotion Lexicon [8] to classify a sentence into one of the categories: anger, joy, sadness, fear, disgust, surprise and neutral. We showed that the predictions of emotion from text using our method were generally close to those obtained by human annotation, with exception of some emotions which also obtained lower agreement between annotators (particularly for disgust, surprise and fear).

We built the expressive voices using the HMM-based speech synthesis method and an audiobook corpus. Since an audiobook contains a wide variety of speaking styles, the challenge was to determine subsets of the corpus that map out the basic emotions considered in this work. We divided the audiobook corpus into subsets representing different emotions by first applying a clustering technique (Self Organising Map) to the speech data and then using the emotion labels of the sentiment analysis to automatically detect the clusters and utterances that best represent each emotion. Results of a preliminary perceptual experiment showed that the textual sentiment analysis does not always correspond to the emotions conveyed in uttered speech. Nevertheless, we assumed that this correspondence was high enough to obtain convincing expressive voices.

2. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie) and EU FP7-METALOGUE project under Grant No. 611073, at Trinity College Dublin, and by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme.

3. References

[1] Chen, L., Gales, M., Braunschweiler, N., Akamine, M. and Knill, K., "Integrated Expression Prediction and Speech Synthesis From Text", IEEE Journal of Selected Topics in Signal Proc., 8(2):323–335, 2014.

[2] Cahn, Janet E. "The generation of a ect in synthesized speech." Journal of the American Voice I/O Society 8 (1990): 1-19.

[3] Schröder, Marc. "Emotional speech synthesis: a review." INTERSPEECH. 2001.

[4] Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. ACL, 2009.

[5] Trilla, Alexandre, and Francesc Alias. "Sentence-based sentiment analysis for expressive text-to-speech." Audio, Speech, and Language Processing, IEEE Transactions on 21.2 2013. pp. 223–233.

[6] Charfuelan, Marcela, and Ingmar Steiner. "Expressive speech synthesis in MARY TTS using audiobook data and emotionML." INTERSPEECH. 2013.

[7] Afli, Haithem et al. "SentiWordsTweet". adapt-invention disclosure form, 2016 (in Press).

[8] Mohammad, Saif M., and Peter D. Turney. NRC Emotion Lexicon. NRC Technical Report, 2013.

FROM HMMS TO DNNS: WHERE DO THE IMPROVEMENTS COME FROM?

Oliver Watts, Gustav Eje Henter, Thomas Merritt, Zhizheng Wu, Simon King

Deep neural networks (DNNs) have recently been the focus of much text-to-speech research as a replacement for decision trees and hidden Markov models (HMMs) in statistical parametric synthesis systems. Performance improvements have been reported; however, the configuration of systems evaluated makes it impossible to judge how much of the improvement is due to the new machine learning methods, and how much is due to other novel aspects of the systems. Specifically, whereas the decision trees in HMM-based systems typically operate at the state-level, and separate trees are used to handle separate acoustic streams, most DNN-based systems are trained to make predictions simultaneously for all streams at the level of the acoustic frame. This paper isolates the influence of three factors (machine learning method; state vs. frame predictions; separate vs. combined stream predictions) by building a continuum of systems along which only a single factor is varied at a time. We find that replacing decision trees with DNNs and moving from state-level to frame-level predictions both significantly improve listeners' naturalness ratings of synthetic speech produced by the systems. No improvement is found to result from switching from separate-stream to combined-stream predictions.

Automatic Analysis of Motivational Interviewing with Diabetes Patients

Xizi Wei, Martin Russell, Peter Jancovic

School of Engineering, University of Birmingham, Birmingham B15 2TT, UK
XXW395@student.bham.ac.uk, M.J.RUSSELL@bham.ac.uk, P.Jancovic@bham.ac.uk

1. Abstract

Motivational Interviewing is a client-centered counseling style to evoke the client's inner motivation to change his or her behavior by helping the client to explore and resolve ambivalence. The therapist must try to achieve and demonstrate empathy with the client, work to highlight the differences between the client's aspirations and his or her current behaviour, avoid argument and direct confrontation, and accommodate rather than directly oppose, the client's resistance to change. To achieve these goals it is essential that the therapist adheres to a set of guidelines. ([1])
Conventional assessment of Motivational Interviewing is based on human analysis of transcriptions of a therapist-client dialogue, to measure its compliance with guidelines, and subjective judgements of paralinguistic factors such as the 'spirit of the conversation' and engagement. These metrics are labour-intensive and very expensive to apply.

An effective alternative to labour-intensive human judgements is automatic analysis of Motivational Interviewing. Human transcriptions of recordings can be replaced by automatic transcription using computer speech recognition. This is the objective of the current research.

Objectives
The objective of the research is to develop technology to initiate, manage, interpret and assess the Motivational Interview. In order to achieve this, the project will focus on the following objectives:

1. Automatic transcription and diarization of Motivational Interviews: to recognize what was said, who said it and when it was said.
2. Linguistic analysis of (automatically) transcribed Motivational Interviews: for example, to measure the extent to which a therapist adheres to the principles of MI, through the use of open, rather than closed, questions.
3. Paralinguistic analysis of Motivational Interviews: to identify the emotional content of the conversation and to measure engagement between the participants.
4. Topic spotting: to track the topic of conversation and monitor whether the conversation stays "on topic".
5. Automatic dialogue modelling: application of statistical models such as POMDPs to model the dialogue structure of Motivational Interviews.
6. Creation of a complete interactive spoken dialogue system to initiate, guide and interpret Motivational Interviews.

Resources
1. IoPPN have a large collection of recordings of Motivational Interviews. Transcriptions of 8 full sessions and 100 10-minute extracts are already available.

2 References

[1] Motivationalinterview.net. (2016). *What is MI?*. [online] Available at: http://www.motivationalinterview.net/clinical/whatismi.html [Accessed 10 Jun. 2016].

Log-linear System Combination Using Structured Support Vector Machines

J. Yang, A. Ragni, M. J. F. Gales and K. M. Knill

Department of Engineering, University of Cambridge
{jy308,ar527,mjfg,kate.knill}@eng.cam.ac.uk

Building high accuracy speech recognition systems with limited language resources is a highly challenging task. Although the use of multi-language data for acoustic models yields improvements, performance is often unsatisfactory with highly limited acoustic training data. In these situations, it is possible to consider using multiple well trained acoustic models and combine the system outputs together. Unfortunately, the computational cost associated with these approaches is high as multiple decoding runs are required. To address this problem, this paper examines schemes based on log-linear score combination. This has a number of advantages over standard combination schemes. Even with limited acoustic training data, it is possible to train, for example, phone-specific combination weights, allowing detailed relationships between the available well trained models to be obtained. To ensure robust parameter estimation, this paper casts log-linear score combination into a structured support vector machine (SSVM) learning task. This yields a method to train model parameters with good generalisation properties. Here the SSVM feature space is a set of scores from well-trained individual systems. The SSVM approach is compared to lattice rescoring and confusion network combination using language packs released within the IARPA Babel program.

In our previous work [1], structured discriminative models are trained using the feature space based on phone log-likelihoods with the same context but different central phone generated by tandem and hybrid systems. Small gains were observed from using additional log-likelihoods extracted from the same models. [2] examines combination of hybrid and tandem systems with log-linear models, and applies learnt phone-specific combination weights to frame level joint decoding, achieving a small performance gain. In this work [3], the log-likelihood score combination approach is investigated. This approach is cast into a structured support vector machine (SSVM) learning task to robustly estimate phone-specific combination weights. A more meaningful feature space is used, which is based on phone log-likelihoods from multiple systems, rather than using extra phone log-likelihoods with the same context extracted from a single system.

References

[1] R. C. van Dalen, J. Yang, H. Wang, A. Ragni, C. Zhang, and M. J. F. Gales, "Structured discriminative models using deep neural-network features," in *Proceedings of ASRU*, 2015.

[2] J. Yang, C. Zhang, A. Ragni, M. J. F. Gales, and P. C. Woodland, "System combination with log-linear models," in *Proceedings of ICASSP*, 2016.

[3] J. Yang, A. Ragni, M. J. F. Gales, and K. M. Knill, "Log-linear system combination using structured support vector machines," to appear in *Proceedings of Interspeech*, 2016.