Fifth Speech Conference of UK and Ireland
held at the University of East Anglia, Norwich UK
2nd-3rd July 2015

We are grateful for the sponsorship of
Microsoft Research

Programme

Thursday July 2nd

13:45–14:00	S0.31	Welcome
14:00–15:00	S0.31	*Computational Paralinguistics: Breaking the Voice*, Björn Schuller, University of Passau and Imperial College, London
15:00–16:00	S0.31	Oral session 1 (see programme)
16:00–16:15	d'Arcy Thompson	Tea
16:15–16:45	S0.31	Poster spotlight talks
16:45–18:00	d'Arcy Thompson	Poster session (see programme)
18:15–19:00	SCVA	Private Viewing of *Bacon and the Masters* exhibition (optional).
19:00–	SCVA	Reception and dinner in the Sainsbury Centre for the Visual Arts, sponsored by Microsoft Research

Friday July 3rd

09:00–10:00	S0.31	*An Overview of HTK v3.5*, Phil Woodland, University of Cambridge
10:00–10:30	S0.31	*What's happening in Audio-visual speech processing?*, led by Barry Theobald, UEA
10:30–11:00	S0.31	*Avatar Therapy: an experience of applying speech technology in healthcare*, Mark Huckvale, UCL
11:00–11:30	d'Arcy Thompson	Coffee
11:30–12:00	S0.31	Poster spotlight talks
12:00–13:30	d'Arcy Thompson	Poster session during lunch (provided) (see programme)
13:30–14:50	S0.31	Oral session 2 (see programme)
14:50–15:00	S0.31	Concluding remarks
15:00–		Depart

Venues (see map):

S0.31	is the lecture theatre next to the entrance to the Schools of Computing Sciences and Mathematics.
d'Arcy Thompson	is on the second floor up the stairs at the entrance to S0.31.
SCVA	is the Sainsbury Centre for Visual Arts, a large art gallery on the West side of the campus. Please use the entrance in the middle of the building.

Editors: Richard Harvey, Stephen Cox, Barry Theobald

Keynote 1

Computational Paralinguistics: Breaking the Voice

Björn Schuller, Chair of Complex & Intelligent Systems, University of Passau, Germany and Machine Learning Group, Imperial College London, United Kingdom

Whether the voice is the actual mirror of our soul or not, it certainly reveals a multitude of aspects about oneself. Hearing a voice on the phone for the first time, one often immediately starts making up one's mind on the age, gender, emotion, personality, likability, origin, and many further characteristics about the counter-part on the other end of the line. Computers are recently starting to go beyond this point, such as when automatically assessing potential autism spectrum condition in new-borns' early vocalisations or the degree of Parkinson's condition from the voice. Here, an overview shall be given on tasks targeted already successfully by the research community. This comes in line with performances of our current-gen automatic voice analysis engines as observed in today's competitive challenges. While these are often impressive, there is an urgent desire to learn more about the interplay and co-influence of the diverse states and traits that ultimately impact on the same vocal production mechanism. Only if this is well understood, one can assume robust characterisation of speakers to work highly independently of side conditions. To this end, more speech data will be needed helping to break the voice code that comes with a rich variety of labels featuring more than one aspect at a time. Likewise, avenues towards efficient exploitation of the vast amount of (unlabelled) speech data available are discussed likely enabling the next generation of Computational Paralinguistics engines. This includes discussion of technical necessities allowing such weakly supervised dynamic cooperative deep and transfer learning from "big data" such as confidence measure estimation and distribution.

Oral Session 1

mark.sinclair@ed.ac.uk

Speech Segmentation and Speaker Diarisation for Transcription and Translation

This dissertation outlines work related to Speech Segmentation -- segmenting an audio recording into regions of speech and non-speech, and Speaker Diarization -– further segmenting those regions into those pertaining to homogeneous speakers.

Knowing not only what was said but also who said it and when, has many useful applications. As well as providing a richer level of transcription for speech, we will show how such knowledge can improve Automatic Speech Recognition (ASR) system performance and can also benefit downstream Natural Language Processing tasks such as Machine Translation and Punctuation Restoration.

While segmentation and diarization may appear to be relatively simple tasks to describe, in practise we find that they are very challenging and are, in general, ill-defined problems. Therefore, we first provide a formalisation of each of the problems as the sub-division of speech within acoustic space and time. Here, we see that the task can become very difficult when we want to partition this domain into our target classes of speakers, whilst avoiding other classes that reside in the same space, such as phonemes. We then examine the tasks in more detail and introduce existing methods and research.

Current Speaker Diarization systems are notoriously sensitive to hyper-parameters and lack robustness across datasets. Therefore, we present a method which uses a series of oracle experiments to expose the limitations of current systems and to which system components these limitations can be attributed. We also demonstrate how Diarization Error Rate (DER), the dominant error metric in literature, is not a comprehensive or reliable indicator of overall performance or of error propagation to subsequent downstream tasks. These results inform our subsequent research.

We find that, as a precursor to Speaker Diarization, the task of Speech Segmentation is a crucial first step in the system chain. Current methods typically do not account for the inherent structure of spoken discourse. As such, we explored a novel method which exploits an utterance-duration prior in order to better model the segment distribution of speech. We show how this method improves not only segmentation, but also the performance of subsequent ASR and Machine Translation systems.

Typical ASR transcriptions do not include punctuation and the task of enriching transcriptions with this information is known as Punctuation Restoration. The benefit is not only improved readability but also better compatibility with NLP systems that expect sentence-like units such as in conventional Machine Translation. We show how segmentation and diarization are related tasks that are able to contribute acoustic information that complements existing linguistically-based punctuation approaches.

Finally, we turn our attention back to the Speaker Diarization task itself. Current systems typically use feature sets that are borrowed from ASR. Such features are not necessarily optimal for speaker discrimination and it is only the enforced duration constraints that prevent the system from partitioning speech into the smaller units these features were designed to recognise -- namely phonemes. We therefore develop a method that uses speaker-discriminative Deep Neural Networks to generate features that are more appropriate to the task.

UKSpeech Conference, University of East Anglia, 2/3 July 2015

Using ASR to get data out, not in

John Yardley[1], David Fox[1], Thomas Michel[1,2], Gordon Hunter[2], James Denholm-Price[2]

[1]JPY Ltd, 5 Surbiton Hill Road, Surbiton, Surrey, KT6 3AX, UK
[2]Faculty of Science, Engineering and Computing, Kingston University, Kingston-upon-Thames, KT1 2EE, UK

Abstract

Most users of computers and hand-held devices see Automatic Speech Recognition (ASR) as a way of getting data in - typically transcribing speech in real-time. In a novel web service called *Threads*, we are working on ways to use ASR to get data out by extracting keywords from 'phone calls for classifying and searching. *Threads* unifies various types of digital message so that an organisation can share its non-confidential communications among authorised users. Since most messages are not confidential, this can produce significant benefits in efficiency and collaboration. Confidential and private messages are filtered out using a number of different techniques which allay the concerns of most organisations. The demands on an ASR system for analysing recorded telephone messages are dramatically lower than for a Dictaphone-type application. Firstly, there is no need for the process to operate in real-time since searching typically takes place days, weeks or months after the 'phone call takes place. Secondly, there is little need to recognise words with low information value (e.g. "and", "but", etc) since these words have similarly low value in terms of searching power. Last, but by no means least, related text-based messages such as emails, can provide a large amount of contextual information which is totally absent in a dictation scenario. This information can lead to a specialised language model for the topic of the call. *Threads* can also be configured to utilise various types of ASR process, according to the users' needs and/or budgets. As better processes become available, Threads can use them. Research is also underway to investigate ways of utilising speaker recognition to identify and authenticate callers unambiguously - for example from switchboard numbers - where caller is likely to be a member of a small, closed set of people. In this "work-in-progress" presentation, the authors discuss the Threads' framework and the interesting application of ASR that Threads presents. Our work utilises not only the several years' worth of operational business messages from the Threads' creators, JPY Ltd, but also the extremely large, public-domain Enron corpus of email and 'phone calls. Indeed, the *Threads* Enron Database (TED: http://www.threads.uk.com/threads-enron-database/) is the believed to be the world's only on-line searchable resource for Enron data.

GOING WIDE - AN APPROACH TO IMPROVING NOISE ROBUSTNESS

JI MING, DANNY CROOKES
QUEEN'S UNIVERSITY BELFAST

Abstract

Most deep neural network (DNN) systems focus on discriminating relatively short speech segments and hence have limited robustness to untrained noise. In this presentation, we propose an alternative approach for speech enhancement by modeling very long speech segments, i.e., going wide, rather than deep. We begin by describing an experiment.

We took a clean speech database (TIMIT) and expressed each training sentence as a short-time power spectrum (STPS) sequence $S = (s_1, s_2, ..., s_T)$, where s_t is the STPS vector at time t. Then we took each core test sentence, added different types of noise (airport, babble, car, restaurant, street, and train station) at an SNR of 0 dB, and converted it to a STPS sequence $X = (x_1, x_2, ..., x_\mathcal{T})$, where each noisy STPS vector x_t can be approximately expressed as $x_t = s'_t + n_t$, with s'_t representing the underlying speech STPS vector and n_t representing the noise STPS vector. For each noisy X, we aimed at finding a clean speech STPS sequence that matches the underlying $(s'_1, s'_2, ..., s'_T)$ from the training data. We obtained an estimate of s'_t by maximizing the following normalized sample correlation coefficient over all the training data

$$\hat{s}_\tau = \arg\max_{s_\tau} R(X_{t\pm L}, S_{\tau\pm L}) = \max_{s_\tau} \frac{\sum_{l=-L}^{L} (x_{t+l} - m_X)^{\mathrm{T}} (s_{\tau+l} - m_S)}{\sigma_X \sigma_S}$$

where $X_{t\pm L}$ denotes a segment of noisy STPS vectors surrounding x_t from x_{t-L} to x_{t+L}, m_X is the mean vector of $X_{t\pm L}$, and σ_X is the Euclidean norm of the mean-removed $X_{t\pm L}$; the same definitions apply to the training speech segment $S_{\tau\pm L}$. *In the experiment, we included the clean test sentence (i.e., the perfect match) in the training data, to examine under what condition it would be chosen.* Fig. 1 shows the accuracy rates of finding the perfect match as a function of the segment length L, averaged over all the times t of all the core test utterances. In theory, it can be shown that as L increases, the matching accuracy will become independent of the noise (it will depend only on the correlation between the speech segments being compared), for the noises that are independent of the speech.

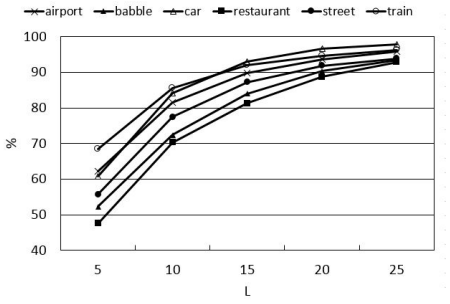


FIGURE 1. Frame identification accuracy as a function of the segment length $2L + 1$.

The above experimental results and the theory behind them suggest the potential of the approach. When the test speech is unseen and noisy, we concatenate a number of short training segments into full sentences (i.e., the longest possible speech segments for the given noisy sentences) with maximum normalized correlation coefficients subject to the independence of the noise, to obtain noise-robust speech estimates. We have tested the new method for single-channel speech enhancement without any estimation of the noise and obtained better performance than other speech enhancement algorithms.

Poster Session 1

Speaker Appeal in Political Speech

Ailbhe Cullen, Andrew Hines, and Naomi Harte

Abstract

Recent years have seen an increase in the volume and diversity of political material available on the internet, including audio and video recordings of parliamentary debates, interviews, and propaganda videos. The challenge for the user is to sort through this data, in order to find something appealing or engaging. In this presentation, we explore the relationship between the vocal characteristics of a politician and their perceived appeal. In particular, we wish to uncover the effect of changing location and audience on the perception of speaker appeal, and on our ability to automatically classify appealing clips.

For the purposes of this study, we have collected a database of Irish political speech. The database contains recordings of a single speaker, the Prime Minister of Ireland, in four situations: interview, election rally, party-political conference, and parliament. This is different to previous political speech corpora which typically contain a single interaction scenario [1]. We choose a single speaker in order to remove variance between speakers, and to allow us to focus on the effects of location, audience, and motivation. Other domains, such as speech recognition, can benefit from speaker specific models. Thus it makes sense to exploit speaker specific traits in a retrieval context, wherever possible. Furthermore, studies within the political science domain are often concerned with the in-depth analysis of a single speaker [2, 3].

Our database has been annotated for six speaker attributes - boring, charismatic, enthusiastic, inspiring, likeable, and persuasive. The annotation of such paralinguistic traits is known to be difficult [4], due to the inherent subjectivity of the task. Thus, we first demonstrate the reliability of the obtained labels. This reliability can be improved by combining the base attributes into a single metric, to form a measure of overall speaker appeal (OSA). In order to demonstrate the benefit of this label aggregation, we perform binary classification, using both the original charisma and derived OSA labels. The performance achieved using the OSA labels is consistently higher than that of the charisma labels. This is in agreement with previous studies which have found that performance increases as the reliability of the labels increases [5, 6].

It is established that the situation, or genre, of political speech affects the speaking style [3, 7]. Similarly, variation has been observed in the prosodic behaviour of speakers in different acoustic environments [8]. Thus, the next question we pose is how the variation of situation and acoustic environment affects the perception of speaker appeal. Despite the findings in [8], there is no significant effect of acoustic environment on charisma or OSA ratings. However, we find a significant bias in the labels due to the situation of the recording.

Finally, having established that the situation influences speaker appeal, we ask the question how can we exploit this knowledge to improve the detection of OSA from the voice. We find that through a combination of aggregating base attributes into a single cover class, and situation specific modelling we can significantly improve classification performance.

References

[1] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, "Predicting continuous conflict perception with bayesian gaussian processes," *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 187 – 200, 2014.

[2] A. Finlayson and J. Martin, "'it ain't what you say...': British political studies and the analysis of speech and rhetoric," *British Politics*, vol. 3, no. 4, pp. 445–464, 2008.

[3] J. W. Pennebaker and T. C. Lay, "Language use and personality during crises: Analyses of mayor rudolph giuliani's press conferences," *Journal of Research in Personality*, vol. 36, no. 3, pp. 271–282, 2002.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] J. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *Multimedia, IEEE Transactions on*, vol. 15, no. 1, pp. 41–55, 2013.

[6] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, "The voice of leadership: Models and performances of automatic analysis in online speeches," *Affective Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 496–508, 2012.

[7] R. B. Slatcher, C. K. Chung, J. W. Pennebaker, and L. D. Stone, "Winning words: Individual differences in linguistic style among u.s. presidential and vice presidential candidates," *Journal of Research in Personality*, vol. 41, no. 1, pp. 63–75, 2007.

[8] A. Astolfi, A. Carullo, L. Pavese, and G. E. Puglisi, "Duration of voicing and silence periods of continuous speech in different acoustic environments," *Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 565 – 579, 2015.

Speech-Based Location Estimation of First Responders in a Simulated Search and Rescue Scenario

Saeid Mokaram , Roger K. Moore

SpandH, Department of Computer Science, University of Sheffield, S1 4DP, United Kingdom
s.mokaram@sheffield.ac.uk, r.k.moore@sheffield.ac.uk

In our research, we explore possible solutions for extracting valuable information about first responders' (FR) location from speech communication channels during crisis response. Fine-grained identification of fundamental units of meaning (e.g. sentences, named entities and dialogue acts) is sensitive to high error rate in automatic transcriptions of noisy speech. However, looking from a topic-based perspective and utilizing text vectorization techniques such as Latent Dirichlet Allocation (LDA) make this more robust to such errors. In this paper, the location estimation problem is framed as a topic segmentation task on FRs' spoken reports about their observations and actions. Identifying the changes in the content of a report over time is an indication that the speaker has moved from one particular location to another. This provides an estimation about the location of the speaker. A goal-oriented human/human conversational speech corpus was collected based on an abstract communication model between FR and task leader during a search process in a simulation environment. Results show the effectiveness of a topic-based approach and especially low sensitivity of the LDA-based method to the highly imperfect automatic transcriptions.

Index Terms: spoken language understanding, speech recognition, topic segmentation, Latent Dirichlet Allocation (LDA), human/human conversation

A Non-Parametric Articulatory-to-Acoustic Conversion System for Silent Speech using Shared Gaussian Process Dynamical Models

Jose A. Gonzalez[*1], Phil D. Green[†1], Roger K. Moore[‡1], Lam A. Cheah[§2], and James M. Gilbert[¶2]

[1]Department of Computer Science, University of Sheffield
[2]School of Engineering, University of Hull

As part of our ongoing work to develop a silent speech interface (SSI) system for post-laryngectomy speech rehabilitation, this work presents a technique for articulatory-to-acoustic conversion using a non-parametric, statistical approach based on shared Gaussian process dynamical models (SGPDMs). In the proposed technique, simultaneous recordings of articulatory and acoustic data are used to learn a mapping between both domains using a SGPDM, which is a non-parametric model providing a shared low-dimensional embedding of the articulatory and acoustic data as well as a dynamic model in the latent space. The learned model is then used for generating an audible speech signal from captured articulatory data. In this work, articulator motion data from the lips and tongue is captured using a technique known as permanent magnet articulography, in which a set of magnets are attached to the articulators and the variations of the magnetic field generated while the user 'speaks' are sensed by a number of magnetic sensors located around the mouth. Preliminary results show that the proposed mapping is able to synthesise high-quality speech from PMA data for certain restricted tasks, but further research is needed before the technique can be applied to a real-life scenario.

[*]j.gonzalez@sheffield.ac.uk
[†]p.green@sheffield.ac.uk
[‡]r.k.moore@sheffield.ac.uk
[§]l.cheah@hull.ac.uk
[¶]j.m.gilbert@hull.ac.uk

Automatic Assessment Of English Learner Pronunciation Using Discriminative Classifiers

Mauro Nicolao, Amy V. Beeston, Thomas Hain

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

This paper presents a novel system for automatic assessment of pronunciation quality of English learner speech, based on deep neural network (DNN) features and phoneme specific discriminative classifiers. DNNs trained on a large corpus of native and non-native learner speech are used to extract phoneme posterior probabilities. Recordings were made via an online learning environment, by mainly Dutch pupils in schools across the Netherlands. A part of the corpus includes mispronunciations annotated at a phonetic level, which allows training of two Gaussian Mixture Models (GMM), representing correct pronunciations and typical error patterns. The likelihood ratio is then obtained for each observed phone.
The method was implemented with the architecture outlined in Figure 1.

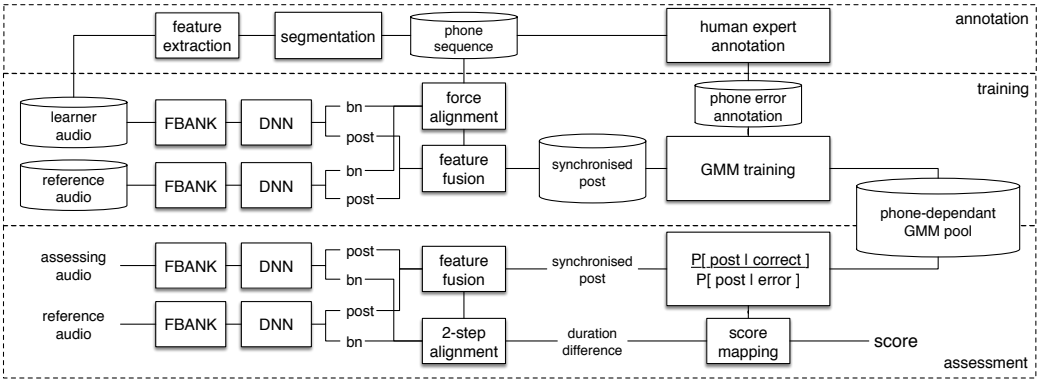


Figure 1: Pronunciation evaluation framework showing stages of annotation (*top*), training (*middle*) and assessment (*bottom*).

During training, learner and teacher audio fragments are fused in a pairwise manner, phone-by-phone. During assessment, a '2-step' process first obtains target phone sequence from the teacher's audio using a multiple-pronunciation dictionary, and secondly uses this sequence in forced alignment of the student's recording. Teacher and learner phones typically have different durations; same-length feature sets are created here by interpolation of the extracted vectors along each dimension. Likelihood ratio value is eventually mapped into the [0,1] interval using a regression tree, a measure of the time-discrepancy between teacher and student phone durations, and their differential values.
Several phone-dependent GMM models were evaluated on a large corpus of English-learning students, with a variety of skill levels, and aged 13 upwards. The cross-correlation of the best system and average human annotator reference scores is 0.72, with miss and false alarm rate around 19%. Automatic assessment is 81.6% correct with a high degree of confidence.
The new approach significantly outperforms our spectral distance based baseline systems.

Automatic Accent Recognition for Forensic Applications

Georgina Brown and Dominic Watt
Department of Language and Linguistic Science, University of York, UK

Previous research on automatic accent recognition has largely involved accents exhibiting great degrees of variation from one another. This task has shown to benefit automatic speech recognition in model adaptation (e.g. Najafian *et al*, 2014). Another application which might benefit from this sort of technology, however, is the forensic application. Forensic analysts might be given the task of extracting as much information about a speaker from a speech recording as possible, and useful distinguishing information is likely to lie in the accent the speaker uses. In more tightly-defined areas in which only minor accent differences exist, speaker profiling may require very fine-grained analysis. This paper assesses the sensitivity of different automatic accent recognition systems by testing them on a corpus of accents which do not differ from one another as markedly as those represented in alternative corpora. For this purpose, we have used the *Accent and Identity on the Scottish/English Border* (AISEB) corpus (Watt *et al*, 2014). The systems being tested are GMM-UBM, GMM-SVM, phonological GMM-SVM, Y-ACCDIST-Correlation and Y-ACCDIST-SVM. The Y-ACCDIST (York-ACCDIST) systems are both based on the ACCDIST metric (Huckvale, 2004), but have been adapted to make processing content-mismatched data possible. Results are shown below:

Table 1: Recognition rates for five accent recognition systems (chance expectation approx. 25%).

System	**% Correct**
GMM-UBM	37.5
GMM-SVM	35.0
Phon-GMM-SVM	65.0
Y-ACCDIST Correlation	82.5
Y-ACCDIST-SVM	87.5

We can observe great differences in performance between the systems. This poster will show system preferences among the individual varieties by focussing on individual confusion matrices. For example, all Eyemouth speakers are correctly classified by both Y-ACCDIST systems, whereas Carlisle speakers are more likely to be correctly classified by Phon-GMM-SVM.

We also demonstrate the contribution that feature selection methods can offer a forensic analysis. Taking the highest-performing system, Y-ACCDIST-SVM, this paper compares Recursive Feature Elimination (RFE-SVM) and ANOVA (applied in Wu *et al* (2010)). While feature selection aims to improve performance, a ranking of speech segments could also be of great value to a forensic analyst when working with unfamiliar accent varieties. We therefore suggest a new way of screening an accent dataset which may be of use to both forensic speech science and sociophonetic research.

References

Huckvale, M. (2004), ACCDIST: A metric for comparing speakers' accents, *in* 'Proceedings of the International Conference on Spoken Language Processing', Korea, pp. 29–32.

Najafian, M., DeMarco, A., Cox, S. & Russell, M. (2014), Unsupervised model selection for recognition of regional accented speech, *in* 'Proceedings fo Interspeech', Singapore, pp. 2967–2971.

Watt, D., Llamas, C. & Johnson, D. E. (2014), Sociolinguistic variation on the Scottish-English border, *in* R. Lawson, ed., 'Sociolinguistics of Scotland', Palgrave Macmillan, London, pp. 79–102.

Wu, T., Duchateau, J., Martens, J.-P. & Compernolle, D. V. (2010), 'Feature subset selection for improved native accent identification', *Speech Communication* **2**, 83–98.

Objective measures for predicting the intelligibility of spectrally smoothed speech with artificial excitation

Danny Websdale, Thomas Le Cornu and Ben Milner
University of East Anglia

Abstract

A study is presented on how well speech quality and intelligibility objective measures can predict the subjective intelligibility of speech that has undergone spectral envelope smoothing and simplification of its excitation. Speech modifications are made by resynthesising speech using the STRAIGHT vocoder, which requires a time-frequency surface (spectral envelope), fundamental frequency and aperiodicity (excitation). Two representations of spectral-envelope (LPC and Filterbank) and four excitation methods (original, monotone, time-varying and unvoiced), are presented and undergo subject intelligibility tests. Objective measures are applied to the modified speech and include measures of speech quality, signal-to-noise ratio, intelligibility and a new proposed measure, the normalised frequency-weighted spectral distortion (NFD) measure. The measures are compared to subjective intelligibility scores using Pearson's correlation, where it is found that several have high correlation ($|r| \geq 0.7$), with NFD achieving the highest correlation ($r = -0.81$).

DNN-Based Missing-Data Mask Estimation for Noise-Robust ASR in Dual-Microphone Smartphones

Iván López-Espejo[*], *José A. González*[†], *Angel M. Gomez*[*], *and Antonio M. Peinado*[*]

[*]Dpt. of Signal Theory, Telematics and Com., University of Granada, Spain
[†]Dpt. of Computer Science, University of Sheffield, UK
{iloes,amgg,amp}@ugr.es, j.gonzalez@sheffield.ac.uk

Automatic speech recognition (ASR) technology is experiencing a new upswing in recent times thanks to the latest portable electronic devices (e.g. smartphones or tablets). In addition, these devices are beginning to integrate small microphone arrays (i.e. microphone arrays composed by a few number of sensors) especially intended to perform noise reduction on the speech signal. While this small microphone array feature is especially being exploited for speech enhancement purposes, few benefit is being taken for noise-robust ASR. Thus, we wish to present some of the advances recently achieved in our research group on the topic of noise-robust ASR with small microphone arrays. In particular, we show that the dual-channel information provided by a smartphone with a dual-microphone can be exploited to easily estimate accurate missing-data masks to perform noise-robust ASR. The followed approach is based on deep neural networks (DNNs), which have demonstrated to be a powerful tool in the field of signal processing in many different ways. Once the missing-data mask is estimated, its quality is evaluated both in terms of the percentage of wrongly estimated mask bins and the word accuracy obtained when used by a spectral reconstruction method. The considered spectral reconstruction method is called truncated-Gaussian based imputation (TGI). Moreover, such experiments are performed on the AURORA2-2C-CT (Aurora-2 - 2 Channels - Close-Talk) database, also developed in our research group. The AURORA2-2C-CT is based on the well-known Aurora-2 database and emulates the acquisition of noisy speech signals with a dual-microphone smartphone used in close-talk conditions (i.e. when the loudspeaker of the smartphone is placed at the ear of the user). Our experimental results show that the DNN is able to exploit the dual-channel information in a simple and efficient way outperforming state-of-the-art single-channel noise-robust approaches.

Deep learning for ASR of children's speech

Mengjie Qian[1], *Ian McLoughlin*[1,2], *Martin Russell*[3]

[1] National Engineering Laboratory of Speech and Language Information Processing, The University of Science and Technology of China, Hefei, Anhui, China.
[2] School of Computing, The University of Kent, Medway Campus, Chatham, UK
[3] School of Electronic, Electrical and Systems Engineering, The University of Birmingham, UK
qmj@mail.ustc.edu.cn, ivm@ustc.edu.cn, m.j.russell@bham.ac.uk

Abstract: Automatic speech recognition (ASR) for children's speech is more difficult than for adults' speech due not only to a larger range of *f0* and formant frequencies than adults, but more severe pauses, false starts, mistakes, non-speech sounds and speed variation than for typical adult speech. As deep neural networks (DNN) perform well in automatic speech recognition, so we use DNN for children's speech recognition and compare the result with adults' speech. Thus, this research explores the ability of DNN-based learning systems to cope with the view variability in speech patterns of children aged from 4 to 14.

Introduction: We first train a context-independent model, and then use this model to train the context-dependent model. When training the context-independent model, we calculate the global mean and variance then split the training dataset, converting word MLF files to phone MLF files without spaces. Next, create mono-phones before adding spaces back into the model and convert word MLF files to phone MLF files with sp, and creating mono-phone with sp to yield the mono-phone model. When we evaluate this for PF-STAR, the result is not particularly good (see Table 1). So, we do force-alignment and use the aligned data to train a DNN model. Then test this we see a substantial improvement**.** In these initial results, the state-of-the-art DNN based system is trained and evaluated separately for TIMIT, WSJ and PF-STAR data [1]. Figures 1 to 3 show the sequence of operations, covering data preparation and training stages, initial mono-phone training, and forced alignment procedures.

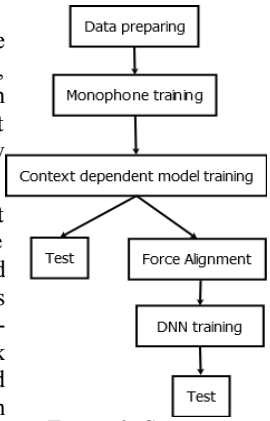


Figure 1: Stages in training and evaluation for different datasets.

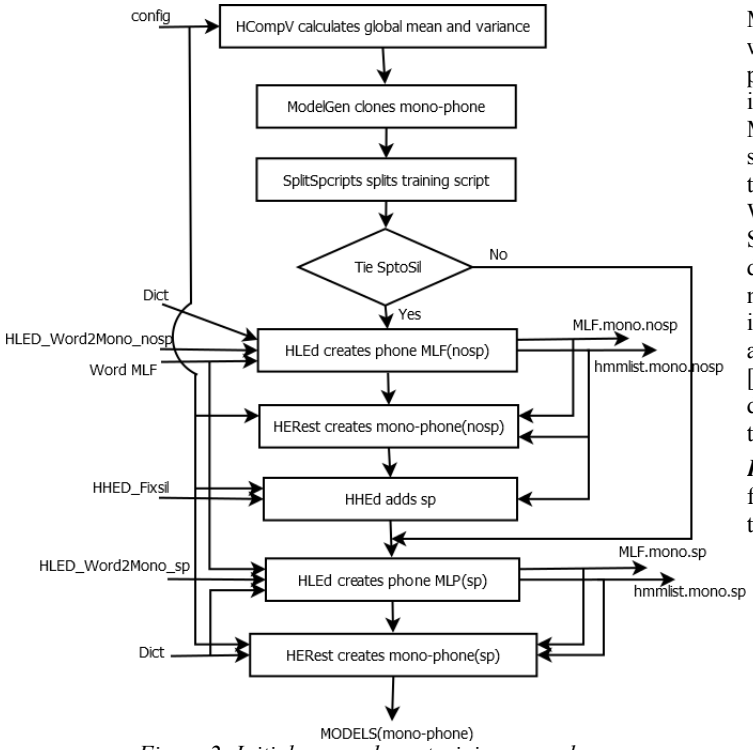


Figure 2: Initial mono-phone training procedure.

Results: are given in Table 1 below in terms of Word Accuracy for three tasks, tested using a baseline GMM-HMM, and with the DNN back end:

Word accuracy	TIMIT	WSJ	PF-STAR
HMM-GMM	71.63	95.54	65.04
DNN	76.92	96.11	82.29

Table 1: Word accuracy results for TIMIT, WSJ and PF-STAR for both systems.

Discussion: the DNN structure yields improvements on TIMIT and WSJ recognition, as has been widely reported by others [2]. However the initial findings are that the DNN is able to provide a much more significant improvement for PF-STAR children's speech database [1]. These results are believed at present to be due to the greater ability of the DNN to extract meaningful discriminative information from higher variance training material. Clearly, this is not the only evaluation of children's speech ASR to have used the DNN structures [3,4]. Reported word accuracies of around 69% [4] can be broken into older and younger age results (with the former performing much better). Further evaluation is currently ongoing.

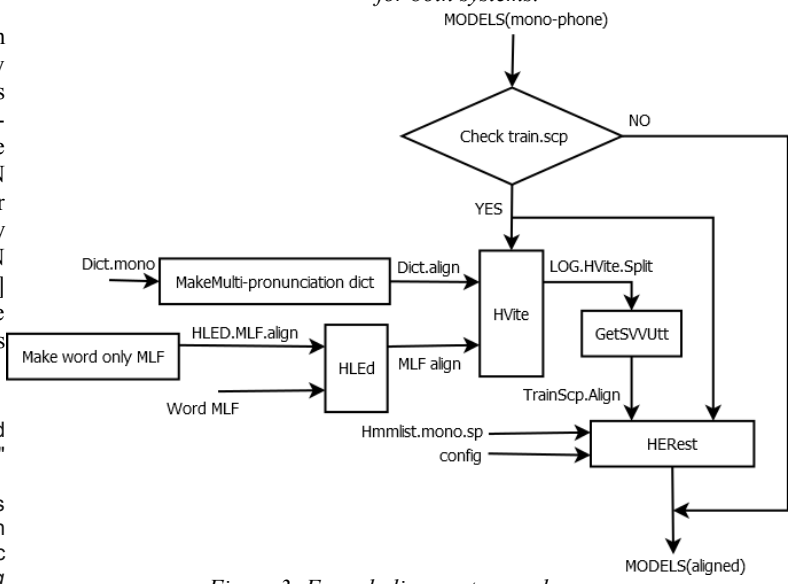


Figure 3: Forced alignment procedure.

[1] Blomberg, Mats, and Daniel Elenius. "Collection and recognition of children's speech in the PF-Star project." *Proc. of Fonetik.* 2003.

[2] Pan, Jia, et al. "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling." *Chinese Spoken Language Processing (ISCSLP), 8th International Symposium on*, 2012.

[3] Serizel, Romain, and Diego Giuliani. "Deep neural network adaptation for children's and adults' speech recognition." *Proc. of the First Italian Computational Linguistics Conference*. 2014.

[4] Elenius, Daniel, and Mats Blomberg. "Comparing speech recognition for adults and children." *Proceedings of FONETIK 2004*: 156-159.

A comparative analysis of continuous and discontinuous implementation of a piecewise linear continuous state HMM.

A conventional HMM approach to speech recognition, relies on a piecewise constant representation of a speech signal. This assumes independence between adjacent features of the data. However, the underlying speech process consists of a continuous, smooth movement of articulators along constrained trajectories. By considering more faithful models of speech that incorporate 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. [4, 5]. 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 [2, 6].

In this work the piecewise linear probabilistic-trajectory segmental model described in [4] is implemented as a special case of the CS-HMM, therefore the compact iterative decoding algorithm described in [2] can be applied instead of the more computationally expensive segmental Viterbi decoder.

The benefit of this implementation is having the freedom to explicitly enforce a continuity constraint between segment boundaries. Consequently, we have a continuous piecewise linear decoder (PLC-Decoder) and a discontinuous piecewise linear decoder (PL-Decoder).

The main focus of this work is to identify whether this continuity constraint improves the piecewise linear segmental model while maintaining a small number of system parameters. For this reason the model currently implements single state phoneme models without any language or complex timing model. A feature representation of the TIMIT data obtained from the bottleneck layer of a neural network [1], comprised of only 9-dimensions is used. A recognition experiment using the full TIMIT test set, yields 53.54 *%Acc* for the PL-Decoder and 50.97 *%Acc* for PLC-Decoder. Although these results are similar, a closer look at the confusion matrices presents some interesting trends.

The PL-Decoder outperforms the PLC-Decoder on the majority of consonant bursts and closures, such as {'b', 'd', 'dx', 'g'…} whereas the PLC-Decoder performs better on *all* of the voiced phonemes. This trend was confirmed by a statistical binomial significance test using the confusion matrices as input and identifying which confusions were significantly greater or less in one system or the other.

The results of this experiment are 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.

To understand the two systems in more detail, we look at regions of speech where the significance test leads us to expect the systems to behave differently. By looking at a graphical representation of the data with the trajectories output from the two systems, the goal is to identify where each decoder tends to produce errors, and whether this can be overcome using a system which is a hybrid of the two piecewise linear decoders.

[1] Linxue Bai, Peter Jančovič, Martin J Russell and Philip Weber. "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics". *[Accepted INTERSPEECH* 2015].

[2] Colin Champion and Steve Houghton. "Application of continuous state Hidden Markov Models to a classical problem in speech recognition". Computer Speech & Language, 2015

[3] John N Holmes, Ignatius G Mattingly, and John N Shearme. "Speech synthesis by rule". Language and speech, 7(3): 127–143, 1964.

[4] Wendy J Holmes and Martin J Russell. "Probabilistic-trajectory segmental HMMs". Computer Speech & Language, 13(1): 3–37, 1999.

[5] Martin J Russell and Philip JB Jackson. "A multiple-level linear/linear segmental HMM with a formant-based intermediate layer". Computer Speech & Language, 19(2): 205–225, 2005.

[6] Philip Weber, Steve Houghton, Colin Champion, Martin Russell, and Peter Jančovič. "Trajectory analysis of speech using continuous state Hidden Markov Models". ICASSP 2014 - Speech and Language Processing (ICASSP2014 - SLTC), Florence, Italy, 2014.

Joint Decoding of Tandem and Hybrid Systems for Improved Keyword Spotting on Low Resource Languages

Haipeng Wang, Anton Ragni, Mark J. F. Gales, Kate M. Knill, Philip C. Woodland, Chao Zhang
Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
{hw443,ar527,mjfg,kate.knill,pcw,cz277}@eng.cam.ac.uk

Abstract

Keyword spotting (KWS) for low-resource languages has drawn increasing attention in recent years. The state-of-the-art KWS systems are based on lattices or Confusion Networks (CN) generated by Automatic Speech Recognition (ASR) systems. It has been shown that considerable KWS gains can be obtained by combining the keyword detection results from different forms of ASR systems, e.g., Tandem and Hybrid systems. This paper investigates an alternative combination scheme for KWS using joint decoding. This scheme treats a Tandem system and a Hybrid system as two separate streams, and makes a linear combination of individual acoustic model log-likelihoods. Joint decoding is more efficient as it requires just a single pass of decoding and a single pass of keyword search. Experiments on six Babel OP2 development languages show that joint decoding is capable of providing consistent gains over each individual system. Moreover, it is possible to efficiently rescore the joint decoding lattices with Tandem or Hybrid acoustic models, and further KWS gains can be obtained by merging the detection posting lists from the joint decoding lattices and rescored lattices.

Applying the Avaya Speech Search Engine to the Mandarin Chinese and Korean languages

Wendy Holmes, Avaya Labs Research

The Avaya Speech Search Engine is a tool for searching audio for occurrences of words and phrases. The emphasis is on real-time searching of large volumes of audio, such as is needed in a typical Contact Centre for example. Ease of deployment to different languages and lack of constraints to words in a dictionary are also important requirements. Hence the underlying technology is not based on transcription, but instead uses what is often referred to as a 'phonetic' approach: incoming audio is efficiently converted into an index of phonetic distances, which can then be searched for matches to a phonetic representation of search terms. This approach does not involve any language model, uses much simpler acoustic models than are needed for accurate transcription, and avoids making early decisions about possible words to recognize.

The speech search engine is designed to be independent of language, and uses data from a 'language pack' for the language in which searches are required. A language pack comprises phonetic and acoustic information about the language, including acoustic models, a pronunciation lexicon and letter-to-sound (LTS) rules so that a pronunciation can be generated for any words that are not in the lexicon. Language packs have previously been developed for various European languages and for two varieties of Arabic. This presentation will describe recent work on developing language packs to support Mandarin Chinese and Korean without changing the general system architecture. (For simplicity of working within the existing architecture, the set of features has not been augmented for Mandarin, and therefore for this first system no attempt has been made to capture the tone properties of the language.)

The main special features of both the Korean and Mandarin language packs concern the character handling and pronunciation generation aspects and, in particular, the letter-to-sound (LTS) prediction module. The LTS module uses decision trees, based on the method described in [1]. Although not state-of-the-art, this framework has been used to train decision trees that give useable accuracy for a variety of languages. For Mandarin however, the huge inventory of characters is such that this approach is not practical. As an alternative, the Unihan database [2] provides Pinyin forms of all Simplified Chinese characters, from which pronunciations can be derived. Korean is rather different as here the syllabic Hangul characters can be decomposed into their constituent Jamo alphabetic form, which has a largely predictable relationship with pronunciation. In this case, it has been possible to obtain accurate LTS prediction by using hand-specified phonological rules based on Jamo characters. Thus it is has proved feasible to represent both the Mandarin and the Korean LTS systems using the same decision tree structure that is normally used for the data-trained LTS trees.

References

[1] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules". In *Proc. ESCA Workshop on Speech Synthesis*, pp. 77--80, Australia, 1998

[2] "Unicode Han Database (Unihan)", http://unicode.org/reports/tr38.

Unsupervised Domain Discovery using Latent Dirichlet Allocation for Acoustic Modelling in Speech Recognition

Mortaza Doulaty, Oscar Saz, Thomas Hain

Speech and Hearing Group, University of Sheffield, Sheffield, UK
{mortaza.doulaty, o.saztorralba, t.hain}@sheffield.ac.uk

Abstract

Speech recognition systems are often highly domain dependent, a fact widely reported in the literature. However the concept of domain is complex and not bound to clear criteria. Hence it is often not evident if data should be considered to be out-of-domain. While both acoustic and language models can be domain specific, work in this paper concentrates on acoustic modelling. We present a novel method to perform unsupervised discovery of domains using Latent Dirichlet Allocation (LDA) modelling. Here a set of hidden domains is assumed to exist in the data, whereby each audio segment can be considered to be a weighted mixture of domain properties. The classification of audio segments into domains allows the creation of domain specific acoustic models for automatic speech recognition. Experiments are conducted on a dataset of diverse speech data covering speech from radio and TV broadcasts, telephone conversations, meetings, lectures and read speech, with a joint training set of 60 hours and a test set of 6 hours. Maximum A Posteriori (MAP) adaptation to LDA based domains was shown to yield relative Word Error Rate (WER) improvements of up to 16% relative, compared to pooled training, and up to 10%, compared with models adapted with human-labelled prior domain knowledge.

The bag–of–words assumption in LDA model does not take the order of words into account. In applying LDA for image processing, there are some variants of the original LDA model, such as Spatial LDA which encodes spatial structure with the visual words. A temporal variant of LDA could better handle the temporal nature of speech and needs to be investigated as a future work. Also applying the current technique on bigger and/or less diverse data set needs to be verified to see what would be the new discovered domains and how they are related to domain adaptation.

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

Deep neural network context embeddings for model selection in rich-context HMM synthesis

Thomas Merritt[1], *Junichi Yamagishi*[1,2], *Zhizheng Wu*[1], *Oliver Watts*[1], *Simon King*[1]
[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]National Institute of Informatics, Japan

1. Introduction

HMM speech synthesis systems offer a flexible and adaptable way to synthesise speech. However the naturalness of these systems is consistently rated below natural speech and unit selection systems as observed in the evaluation results from numerous Blizzard Challenges over many years. Many explanations have been given for the causes of this however few formal investigations have been performed.

In previous work, we did formally investigate several hypotheses, including the effects of: over-smoothing of the spectral envelope as a result of averaging over multiple speech samples from differing contexts [1], temporally over-smoothed parameter trajectories as a result of maximum-likelihood parameter generation (MLPG) [2, 3, 1], parameter generation with poor global variance [2, 3, 1], vocoding[4, 3, 1] and independant modelling of parameter streams[4, 3], among others.

The most striking finding of these investigations was that temporal over-smoothing does not have as strong effect on speech quality as was previously believed, but rather that the gap between standard HMM speech synthesis systems and vocoded speech might be significantly closed by avoiding the averaging of speech samples across differing linguistic contexts.

2. Prior work

Rich-context statistical parametric speech synthesis systems are a notable example which aims to remove the effects of across-context averaging [5]. The term 'rich-context' refers to models which are trained only on samples where the context matches exactly and therefore avoids averaging across differing contexts. The primary example is [5], in which Gaussian mean values are calculated within each unique context found in the training data, with variance values being tied in the usual way.

The system introduced in [5] uses the distribution (i.e., Gaussian) selected by the standard tied decision tree as a reference. It then finds the closest untied rich-context model (from a pre-selected subset of all possible models) to this reference, by computing the divergence between the reference distribution and each of the rich context models. This is counter-intuitive. As we know from [1], this tied model is known to be of poor quality as a result of averaging across different contexts and therefore would seem to be a poor reference for rich-context model selection. The whole point of using rich context models is to get away from the tied model, not to find a model that is as close as possible to it.

3. Proposed bottleneck-driven system

Our proposed system is inspired by [5], however does not use the tied model as a reference for rich context model selection. Instead, it performs selection using an acoustically-supervised embedding of the linguistic context, which we derive from the bottleneck layer of a Deep Neural Network (DNN) speech synthesis model [6].

Each unique input to the DNN (i.e., each unique linguistic context) leads to a particular bottleneck feature vector. That is, we can derive a compact vector-space representation of any linguistic context, including those not seen in the training data. We use distance in this vector space as the way to select rich-context models at synthesis time. The DNN-derived embedding is essentially a compression of the linguistic features, but importantly one that has been learned in conjunction with predicting the acoustics. So, for example, acoustically-irrelevant linguistic features will be ignored, and other features will be 'de-noised' and de-correlated.

Table 1: *Conditions included in listening test*

ID	Description	Postfilter
N	Natural speech	n/a
V	Vocoded speech	n/a
D	Stacked bottleneck DNN system [6]	PF
H	Standard tied HMM speech (HTS demo)	GV
F	HMM speech w/ fully untied tree (MDL = 0) – variance parameters from system H	PF
CT	Rich context system [5] – tri-phone pre-selection	PF
CB	Rich context system [5] – bi-phone pre-selection	PF
E	Proposed system w/ Euclidean distance	PF
KL	Proposed system w/ KLD	PF
ETS	Proposed system w/ Euclidean distance – source parameters from system H	PF
KLTS	Proposed system w/ KLD – source parameters from system H	PF

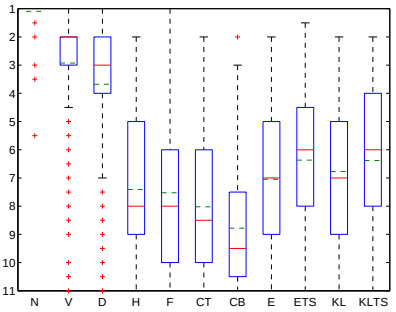


Figure 1: *Boxplot of rank order of conditions from MUSHRA test*

Various measures could be used, at synthesis time, to find the closest rich-context model (in bottleneck feature space) for an unseen context. Here we present two possibilities: Euclidean distance and KLD.

4. Conclusions and future work

The proposed system provides clear improvements on both standard tied HMM models and the previously proposed rich context model system [5]. Although a state-of-the-art DNN setup is better than all HMM systems here, there is further room for improvement in the HMM systems. For example, embeddings could be derived from a DNN that no longer needs to output speech parameters, but perhaps uses more perceptually relevant output features.

The HMM paradigm is much more transparent that the DNN paradigm. Rich-context model parameters can be related directly back to frames in the training data. This link to the training data also suggest simple and obvious ways to build hybrid systems (i.e., statistical model-guided concatenation).

5. References

[1] T. Merritt, J. Latorre, and S. King, "Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech," in *Proc. ICASSP*, 2015.

[2] T. Merritt and S. King, "Investigating the shortcomings of HMM synthesis," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013, pp. 165–170.

[3] T. Merritt, T. Raitio, and S. King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," in *Proc. Interspeech*, 2014, pp. 1509–1513.

[4] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, 2014, pp. 1504–1508.

[5] Z.-J. Yan, Y. Qian, and F. K. Soong, "Rich context modeling for high quality HMM-based TTS," in *Proc. Interspeech*, 2009, pp. 1755–1758.

[6] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *ICASSP*, 2015.

Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features

Zhizheng Wu, Simon King
The Centre for Speech Technology Research, University of Edinburgh

Recently, Deep Neural Networks (DNNs) have shown promise as an acoustic model for statistical parametric speech synthesis. Their ability to learn complex mappings from linguistic features to acoustic features has advanced the naturalness of synthesis speech significantly. However, because DNN parameter estimation methods typically attempt to minimise the mean squared error of each individual frame in the training data, the dynamic and continuous nature of speech parameters is neglected. In this paper, we propose a training criterion that minimises speech parameter trajectory errors, and so takes dynamic constraints from a wide acoustic context into account during training. We combine this novel training criterion with our previously proposed stacked bottleneck features, which provide wide linguistic context. Both objective and subjective evaluation results confirm the effectiveness of the proposed training criterion for improving model accuracy and naturalness of synthesised speech.

1 Relation to prior work

Recently, following the success of Deep Neural Networks (DNNs) as an acoustic model in automatic speech recognition, neural networks have re-emerged as an alternative acoustic model for SPSS, and several studies have presented state-of-the-art performance using DNNs. However, current implementations of DNN-based speech synthesis make a *frame-by-frame independence* assumption during modelling and generation. The frame-by-frame independence assumption has two consequences. The first is the frame-wise independence assumption when predicting acoustic features. Even though contextual information has been included in the linguistic features, when predicting acoustic features for consecutive frames, each frame is generated conditionally independently of the others, given the linguistic context. This has implications for the trajectory of the generated acoustic features. The second consequence is the ignorance of the interaction between static and dynamic features during training DNN models. Dynamic features are extracted from a sequence of static features; but, after this extraction, the relationships between static and dynamic features is neglected during training. This has implications for the accuracy of the acoustic model (i.e., DNN) itself.

Here, we propose a novel training criterion – minimum trajectory error for DNNs – and we combine this with stacked bottleneck features. Rather than minimising frame-wise mean squared error, the minimum trajectory error criterion considers the dynamic feature constraints in the training phase. By integrating this criterion with stacked bottleneck features (which can be viewed as an acoustically-supervised compression and denoising of linguistic context [1]), we can now include contextual constraints at the input linguistic level and the output acoustic level.

2 Proposed minimum trajectory error training

To model the interaction between static and dynamic features and include temporal constraints in the training phase, we propose a new training criterion: to minimise the utterance-level trajectory error, rather than the frame-by-frame error. In this way, we minimise the error of the final smoothed trajectory directly, rather than the intermediate features. In other words, we minimise the error of the output of MLPG (which will of course then be used directly to generate speech), rather than minimising the error of the features that are input to MLPG.

The trajectory error function is defined as,

$$D(\hat{\mathbf{C}},\mathbf{C}) = (\hat{\mathbf{C}}-\mathbf{C})^{\top}(\hat{\mathbf{C}}-\mathbf{C}) \quad (1)$$

$$= (\mathbf{R}\hat{\mathbf{O}}-\mathbf{C})^{\top}(\mathbf{R}\hat{\mathbf{O}}-\mathbf{C}), \quad (2)$$

where $\mathbf{C}$ and $\hat{\mathbf{C}}$ are the reference and generated parameter trajectories, respectively, and $\mathbf{R} = (\mathbf{W}^{\top}\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^{\top}\mathbf{U}^{-1}$ is the matrix to perform parameter generation, given static and delta features. The new error function is computed from the smoothed trajectory rather than the direct output of the DNN. That is, the neural network will model parameter trajectories directly, and hence we need to take the MLPG algorithm into account whilst training the network.

Similar to conventional DNN, gradient descent methods can be used to train the network. The gradients of DNN model parameters λ can computed as:

$$\frac{\partial D(\hat{\mathbf{C}},\mathbf{C})}{\partial \lambda} = \frac{\partial D(\hat{\mathbf{C}},\mathbf{C})}{\partial \hat{\mathbf{O}}}\frac{\partial \hat{\mathbf{O}}}{\partial \lambda} \quad (3)$$

$$= \frac{\partial D(\mathbf{R}\mathbf{O},\mathbf{C})}{\partial \hat{\mathbf{O}}}\frac{\partial \hat{\mathbf{O}}}{\partial \lambda}, \quad (4)$$

where

$$\frac{\partial D(\mathbf{R}\hat{\mathbf{O}},\mathbf{C})}{\partial \hat{\mathbf{O}}} = (\hat{\mathbf{C}}-\mathbf{C})^{\top}\mathbf{R} \quad (5)$$

Here only $\frac{\partial \hat{\mathbf{O}}}{\partial \lambda}$ is directly related to the model parameters. The only difference between the new training criterion and the conventional frame-based mean squared error criterion is the method for computing the errors to be back-propagated through the network. The method for computing gradients for DNN parameters in lower layers is not changed.

3 Experimental results

The preference results are presented in Fig. 3. First, let us examine the effectiveness of the proposed MTE training criterion. It can be observed that MTE-DNN is significantly better than FE-DNN. MTE-BN-DNN also achieves a slightly a higher preference score than FE-BN-DNN, although the difference is not significant.

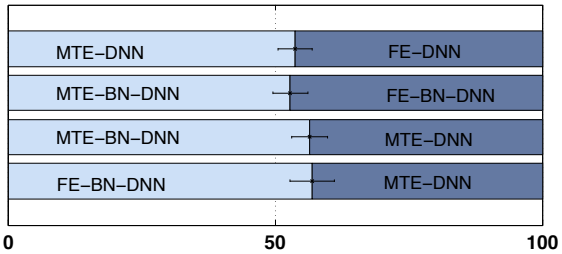


Figure 1: Preference test results for naturalness.

We can also compare the two ways of including temporal constraints by comparing with MTE-DNN and FE-BN-DNN. FE-BN-DNN is significantly better than MTE-DNN in terms of naturalness. This indicates that stacking bottleneck feature at the input level is more effective than considering only temporal constraints at the output acoustic feature level.

Last, we assessed whether the integration of minimum trajectory error criteria and stacked bottleneck features is effective. MTE-BN-DNN is significantly better than MTE-DNN that does not have stacked bottleneck features, and has a slightly (but not significantly) higher listener preference than FE-BN-DNN, which does not use minimum trajectory error criteria. It appears that the minimum trajectory error criterion and stacked bottleneck features approaches are complementary.

4 References

[1] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

Parameterised Sigmoid and ReLU Hidden Activation Functions for DNN Acoustic Modelling

C. Zhang & P. C. Woodland

The form of hidden activation functions has been always an important issue in deep neural network (DNN) design. The most common choices for acoustic modelling are the standard Sigmoid and rectified linear unit (ReLU), which are normally used with fixed function shapes and no adaptive parameters. Recently, there have been several papers that have studied the use of parameterised activation functions for both computer vision and speaker adaptation tasks. In this paper, we investigate generalised forms of both Sigmoid and ReLU with learnable parameters, as well as their integration with the standard DNN acoustic model training process. Experiments using conversational telephone speech (CTS) Mandarin data, result in an average of 3.4% and 2.0% relative word error rate (WER) reduction with Sigmoid and ReLU parameterisations.

SCALING RECURRENT NEURAL NETWORK LANGUAGE MODELS

Will Williams, Niranjani Prasad, David Mrva, Tom Ash, Tony Robinson

Cantab Research, Cambridge, UK

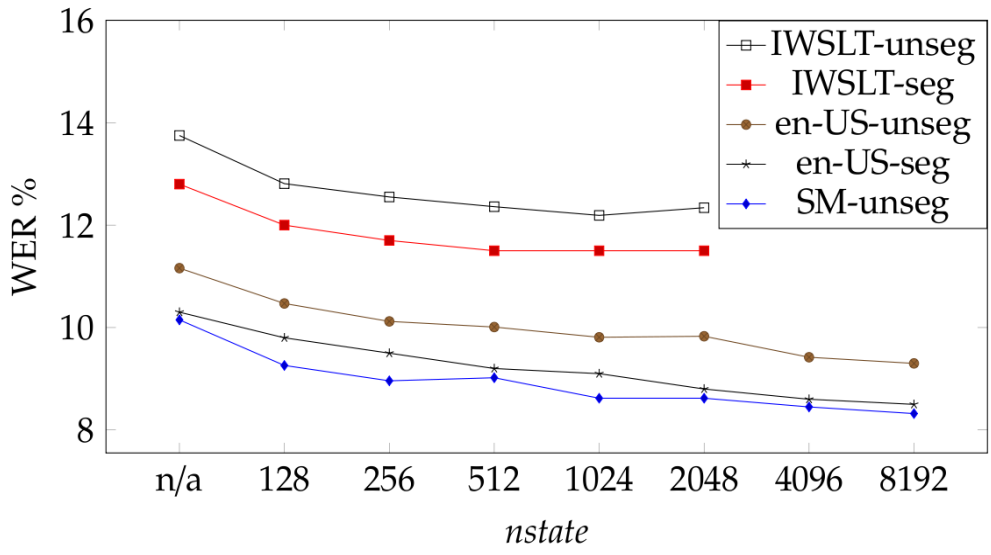
Full paper: http://arxiv.org/abs/1502.00512

This paper investigates the scaling properties of Recurrent Neural Network Language Models (RNNLMs). We discuss how to train very large RNNs on GPUs and address the questions of how RNNLMs scale with respect to model size, training-set size, computational costs and memory.

For high throughput and utilisation on GPUs, we train a standard RNN with stochastic gradient descent and rmsprop, on an internally collated and processed training corpus of 8 billion words.

Our analysis shows that despite being more costly to train, RNNLMs obtain much lower perplexities on standard benchmarks than n-gram models. Although training time is much larger for an RNN and we have to increase the number of hidden state units (nstate) when scaling the training set size, RNNs make much better use of the additional data than n-grams and use far fewer parameters to do so.

We train the largest known RNNLMs and evaluate them by rescoring lattices on three different Kaldi-based ASR systems, using the IWSLT14.SLT.tst2010 test data. Lattices were rescored using a highly efficient internal lattice rescoring tool which operates in considerably less than real time. With both our internal ‘en-US’ system and ‘SM’ (the commercial service available at speechmatics.com) we see an average reduction in WER of 18% relative to rescoring with the n-gram alone by using nstate 8192 RNNLMs.



We also present the new lowest perplexities on the recently released billion word language modelling benchmark, 1 BLEU point gain on machine translation and a 17% relative hit rate gain in word prediction.

An Investigation into Speaker Informed DNN Front-end for LVCSR

Yulan Liu (acp12yl@sheffield.ac.uk), Speech and Hearing (SpandH) group, Department of Computer Science, The University of Sheffield, UK.

Penny Karanasou, Speech Research Group of the Machine Intelligence Laboratory, University of Cambridge, UK.

Thomas Hain, Speech and Hearing (SpandH) group, Department of Computer Science, The University of Sheffield, UK.

Abstract:

Deep Neural Network (DNN) has become a standard method in many ASR tasks. Recently there is considerable interest in "informed training" of DNNs, where DNN input is augmented with auxiliary codes, such as i-vectors, speaker codes, speaker separation bottleneck (SSBN) features, etc. This paper compares different speaker informed DNN training methods in LVCSR task. We discuss mathematical equivalence between speaker informed DNN training and "bias adaptation" which uses speaker dependent biases, and give detailed analysis on influential factors such as dimension, discrimination and stability of auxiliary codes. The analysis is supported by experiments on a meeting recognition task using bottleneck feature based system. Results show that i-vector based adaptation is also effective in bottleneck feature based system (not just hybrid systems). However all tested methods show poor generalisation to unseen speakers. We introduce a system based on speaker classification followed by speaker adaptation of biases, which yields equivalent performance to an i-vector based system with 10.4% relative improvement over baseline on seen speakers. The new approach can serve as a fast alternative especially for short utterances.

Notes:

The same work has been presented as an oral lecture in ICASSP 2015, Brisbane, Australia, in April, 2015.

HMM-based Visual Speech Synthesis using Dynamic Visemes

Ausdang Thangthai and Barry-John Theobald

Speech, Language and Audio Processing Laboratory
{A.Thangthai,B.Theobald}@uea.ac.uk

Abstract

In this paper we incorporate *dynamic visemes* into hidden Markov model (HMM)-based visual speech synthesis. Dynamic visemes represent intuitive visual gestures identified automatically by clustering purely visual speech parameters [Taylor et al., 2012]. They have the advantage of spanning multiple phones and so they capture the effects of visual coarticulation explicitly within the unit. The previous application of dynamic visemes to synthesis used a sample-based approach, where cluster centroids were concatenated to form parameter trajectories corresponding to novel visual speech. In this paper we generalize the use of these units to create more flexible and dynamic animation using a HMM-based synthesis framework. We show using objective and subjective testing with LIPS [Theobald et al., 2008] and KB-2K large audiovisual speech corpus [Taylor et al., 2012]. It shows that a HMM synthesizer trained using dynamic visemes can generate better visual speech than HMM synthesisers trained using either phone or traditional viseme units.
Index Terms: visual speech synthesis, hidden Markov model, dynamic visemes

References

[Taylor et al., 2012] Taylor, S. L., Mahler, M., Theobald, B.-J., and Matthews, I. (2012). Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '12, pages 275–284, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

[Theobald et al., 2008] Theobald, B., Fagel, S., Elsei, F., and Bailly, G. (2008). Lips2008: Visual speech synthesis challenge. In *Proceedings of Interspeech*, pages 1875–1878.

Keynote 2

An Overview of HTK v3.5

Phil Woodland

Machine Intelligence Laboratory
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ

HTK (http://htk.eng.cam.ac.uk) is a research source code toolkit designed primarily for automatic speech recognition (ASR) with more than 100,000 registered users worldwide. HTK v3.5, due to be released later in 2015, will include a number of major new features.

HTK has allowed the development of state-of-the-art hidden Markov model (HMM) speech recognition systems for many years based on Gaussian mixture models (GMMs). However, to include Deep Neural Network (DNN) models, users needed to rely on external tools. HTK v3.5 includes support for artificial neural networks (ANNs) with very general feed-forward architectures to be used for either acoustic modelling or feature extraction. The implementation allows efficient training by supporting GPUs. The ANN modules are fully integrated into the rest of the HTK toolkit to support, for example, sequence training.

HTK v3.5 will also include a new language model interface which enables efficient lattice re-scoring with neural network language models (NNLMs) including recurrent NNLMs. A number of other features include full 64 bit support to handle large data sizes, the use of new dictionary phone attribute markers and state-root decision trees.

Finally, HTK will continue to include comprehensive documentation and examples.

This talk will give more technical details of what is included in HTK v3.5 as well as illustrations of recent speech recognition systems that have been built at Cambridge using this software.

What's happening in audio-visual speech?

Barry-John Theobald, Kwanchiva Thangthai, Richard Harvey, Yuxuan Lan, Stephen Cox, University of East Anglia

In this session we consider recent progress in audiovisual speech processing, with a particular focus on audiovisual speech recognition. The introduction of deep learning has dramatically improved the accuracy of ASR systems that exploit visual information, and we show that a visual- only system (pure lip-reading) can now achieve accuracy approaching 85%.
Many of the restrictions that once were assumed to be required for accurate (visual) recognition of speech, namely well lit faces in lab conditions, minimal head pose variation, accurate feature point tracking, and so on, are less of a problem that once thought. We will outline work that we have been doing in the Audio, Speech and Language Group in the School of Computing Sciences at UEA in collaboration with colleagues at the Centre for Vision, Speech and Signal Processing at the University of Surrey.

Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs

Kwanchiva Thangthai, Richard Harvey, Stephen Cox, Barry-John Theobald

School of Computing Science
University of East Anglia, Norwich, UK
k.thangthai@uea.ac.uk, r.w.harvey@uea.ac.uk, s.j.cox@uea.ac.uk,
b.theobald@uea.ac.uk

Abstract

This work presents preliminary experiments using the Kaldi toolkit to investigate audiovisual speech recognition (AVSR) in noisy environments using deep neural networks (DNNs). In particular we use a single-speaker large vocabulary, continuous audiovisual speech corpus to compare the performance of visual-only, audio-only and audiovisual speech recognition. The models trained using the Kaldi toolkit are compared with the performance of models trained using conventional hidden Markov models (HMMs). In addition, we compare the performance of a speech recognizer both with and without visual features over nine different SNR levels of babble noise ranging from 20dB down to -20dB. The results show that the DNN outperforms conventional HMMs in all experimental conditions, especially for the lip-reading only system, which achieves 84.67% word accuracy (37.19% absolute gain). Moreover, the DNN provides an effective improvement of 10 and 12dB SNR respectively for both the single modal and bimodal speech recognition systems. However, integrating the visual features using simple feature fusion is only effective in SNRs at 5dB and above. Below this the degradion in accuracy of an audiovisual system is similar to the audio only recognizer.

Keynote 3

Avatar Therapy: an experience of applying speech technology in healthcare

Mark Huckvale
Speech, Hearing and Phonetic Sciences
UCL

The hearing of voices is a commonly reported symptom of schizophrenia which can lead to significant problems for sufferers. The voices are sometimes called hallucinations, but they are distressingly real to patients and can impact their ability to have normal social relationships with their family and caregivers or to hold down employment. It is said that 30% of people with this diagnosis continue to experience hallucinations and delusions despite treatment with antipsychotic medication. Auditory hallucinations manifest in a number of ways, including voices that speak aloud what the patient is thinking; voices that give a running commentary on the patient's actions or external imagined events; two or more persons conversing about the patient in the third person; or commands ordering the patient to perform certain actions (which may be violent).

In 2009 Julian Leff proposed that hearing voices sufferers might benefit from engaging in a dialogue with their voices and proposed that we construct an avatar that would represent and embody their persecutory voices. The avatar system that Julian Leff, Geoff Williams and I constructed consisted of a 3D talking head with a customisable appearance and a real-time voice conversion system with a customisable voice transform. Patients first construct an avatar with a face and voice that fits their persecutory voice, then over a small number of therapy sessions with a clinician learn to engage in a dialogue with the avatar and so gain some control over their symptoms. The results of a pilot study were remarkable, with patients reporting a drop in psychotic symptoms and with three patients out of 16 losing their voice hallucinations completely. We are now running a large clinical trial of the technique.

In this talk I will present a personal history of my involvement in Avatar Therapy with a focus on the design challenges and the solutions we adopted. Avatar Therapy is an interesting application for speech technology in that its effects can be life changing, and the problems we faced were not always the ones we expected.

Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., & Leff, A. P. (2013). Silencing voices: a proof-of-concept study of computer-assisted therapy for medication-resistant auditory hallucinations. British Journal of Psychiatry.

Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., & Leff, A. P. (2013). Avatar Therapy for persecutory auditory hallucinations: What is it and how does it work?. Psychosis: Psychological, Social and Integrative Approaches.

Huckvale, M., Leff, J., & Williams, G., (2013) Avatar Therapy: an audio-visual dialogue system for treating auditory hallucinations, Interspeech 2013, Lyon, France.

Poster Session 2

CPrAN: a package manager for Praat

José Joaquín Atria
University College London

June 22, 2015

In the past decade, Praat has become one of the major research tools in use in phonetics and its related fields. This success is due in part to it being Free Software, but also to the availability of a widely used and very versatile scripting language. Paired with its large number of users, this has resulted in a large number of user-generated scripts that are avidly shared, modified, and extended by the community to facilitate research and its reproducibility.

However, the code sharing that does take place does so in a very de-centralised way. Most scripts are written with little to no consideration for code re-use, and when researchers find those scripts they will often modify them to suit other highly-specified tasks of their own. Over time, this has resulted in the development of a fragmented community of users in which similar tools get re-written over and over again, and which in turn gives little incentive for users to write tools designed to suit broader needs. Today, this vicious circle is in full force.

The problem of distribution has seen many attempts at a solution. These will often take the shape of online repositories of scripts, normally maintained by individual researchers who take on the role of managing the entire set of scripts distributed through that platform.

This sort of approach, while popular, is normally unsustainable. Many repositories like these have closed down, and the ones that don't, find it difficult to keep their scripts up to date. They also often don't have the capacity to test the scripts that are distributed, nor do they commonly provide documentation or support. And without lists of available repositories, even keeping track of which of these exist at any one point is difficult for the common user.

An entirely different problem is that of code re-use. The main problem in this case is that code needs to be distributed in a way that is usable by others not only in the direct solution of their immediate needs, but also in the development of their own pieces of code that others might use in the future. Without some means of resolving dependencies, this problem remains unsolved even for those repositories that manage to persist.

The traditional solution for these problems is a package manager, like the ones that exist for Perl (CPAN), for R (CRAN) and for TEX and friends (CTAN). They provide an interface to search, install, remove and update existing packages, manage their dependencies, and often provide facilities for bug reporting and issue tracking.

CPrAN (the Comprehensive Praat Archive Network) aims at providing the Praat community with these same features, to help in the sharing of code, and promote a higher standard of quality in the code that is shared. Today, CPrAN is both an architecture for distribution and a client implementing that architecture. This presentation will cover its design principles as well as the state of the current prototype.

Using Enhanced Videos in Speech Perception Training

Najwa Alghamdi, PhD student.

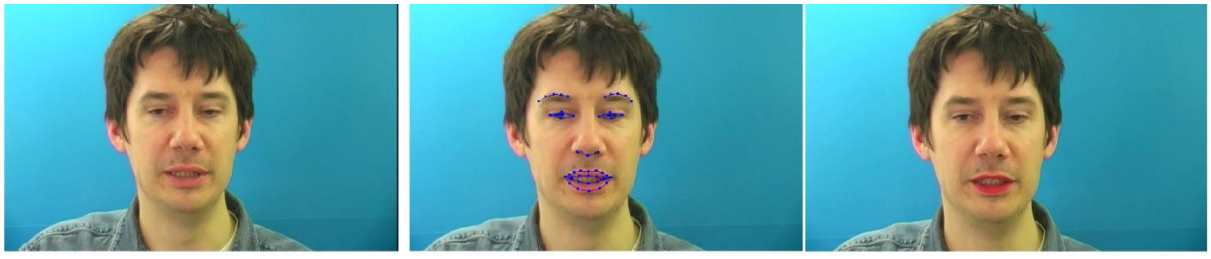
Supervised by Dr. Steve Maddock, Prof. Guy J. Brown and Dr. Jon Barker.

The Department of Computer Science, University of Sheffield.

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 those who use cochlear implants (CI). Previous work suggests that audiovisual training is effective at enhancing hearing abilities for subsequent audio-only situations. This research project investigates the use of artificial enhancement of a speaker's lips in the videos used for such audiovisual training. The first enhancement chosen is to improve the visibility of the lips by automatically tracking them in videos and then artificially colouring them, to simulate the speaker wearing lipstick (Figure 1).

An initial study was carried out using 46 non-native, normal hearing listeners, recruited at King Saud University, Saudi Arabia. CI-simulated speech was used in the training process. The aim was to simulate the experience of real CI users – both non-native listeners and CI users deal with adversity when listening, and thus the non-native listeners may be predictors of the performance of CI users. Three groups were trained with different stimuli – audio-only, audiovisual, or enhanced audiovisual – and audio-only pre- and post-tests were conducted. The results show that the benefit on speech intelligibly gained from introducing a visual signal is increased when the visual signal is enhanced and that the enhanced audiovisual group attained the highest training gain.



(a) Original lips (b) Automatic tracking of the lips (c) Artificial colouring (lip stick) applied to the tracked lips

Figure 1- Enhancing the lips by simulating the speaker wearing a lip stick

Email : amalghamdi1@sheffield.ac.uk

Preferred presentation format: two.

Automatically Grading Learners' Spoken English Using a Gaussian Process

Rogier C. van Dalen, Kate M. Knill, Mark J. F. Gales
ALTA Institute / Department of Engineering
University of Cambridge

June 22, 2015

There is a high demand around the world for the learning of English as a second language. Correspondingly, there is a need to assess the proficiency level of learners both during their studies and for formal qualifications. A number of automatic methods have been proposed to help meet this demand with varying degrees of success. This presentation will consider the automatic assessment of spoken English proficiency, which is still a challenging problem. In this scenario, the grader should be able to accurately assess the learner's ability level from spontaneous, prompted, speech, independent of L1 language and the quality of the audio recording. Automatic graders are potentially more consistent than humans. However, the validity of the predicted grade varies. This presentation will propose an automatic grader based on a Gaussian process. The advantage of using a Gaussian process is that as well as predicting a grade, it provides a measure of the uncertainty of its prediction. The uncertainty measure is sufficiently accurate to decide which automatic grades should be re-graded by humans. It can also be used to determine which candidates are hard to grade for humans and therefore need expert grading. Performance of the automatic grader is shown to be close to human graders on real candidate entries. Interpolation of human and Gaussian process grades further boosts performance.

ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge

Zhizheng Wu[1], Tomi Kinnunen[2], Nicholas Evans[3], Junichi Yamagishi[1], Cemal Hanilçi[2], Md Sahidullah[2], Aleksandr Sizov[2]
[1]University of Edinburgh, United Kingdom [2]University of Eastern Finland, Finland [3]EURECOM, France

An increasing number of independent studies have confirmed the vulnerability of automatic speaker verification (ASV) technology to spoofing. However, in comparison to that involving other biometric modalities, spoofing and countermeasure research for ASV is still in its infancy. A current barrier to progress is the lack of standards which impedes the comparison of results generated by different researchers. The ASVspoof initiative aims to overcome this bottleneck through the provision of standard corpora, protocols and metrics to support a common evaluation. This paper introduces the first edition, summaries the results and discusses directions for future challenges and research [1].

1 Introduction

Automatic speaker verification (ASV) offers a low-cost and flexible biometric solution to person authentication. While the reliability of ASV systems is now considered sufficient to support mass-market adoption, there are concerns that the technology is vulnerable to spoofing, also referred to as presentation attacks. Spoofing refers to an attack whereby a fraudster attempts to manipulate a biometric system by masquerading as another, enrolled person. Acknowledged vulnerabilities include attacks through impersonation, replay, speech synthesis and voice conversion [1].

The ASVspoof challenge aims to encourage further progress through (i) the collection and distribution of a standard dataset with varying spoofing attacks implemented with multiple, diverse algorithms and (ii) a series of competitive evaluations. Following on from the special session in Spoofing and Countermeasures for Automatic Speaker Verification held during the 2013 edition of INTERSPEECH in Lyon, France, the first ASVspoof challeng[2] will be held during the 2015 edition of INTERSPEECH in Dresden, Germany. The challenge has been designed to support, for the first time, independent assessments of vulnerabilities to spoofing and of countermeasure performance. The initiative provides a level playing field to facilitate the comparison of different spoofing countermeasures on a common dataset, with standard protocols and metrics. While preventing as much as possible the inappropriate use of prior knowledge, the challenge also aims to stimulate the development of generalised countermeasures with potential to detect varying and unforeseen spoofing attacks.

This paper describes the ASVspoof database, protocol and metrics, all of which are now in the public domain. Also presented is a summary of 16 sets of participant results. Finally, observations and findings are presented with priorities for the future.

2 Challenge database

ASVspoof is based upon a standard database consisting of both genuine and spoofed speech. Genuine speech is recorded from 106 human speakers (45 male and 61 female) without any modification, and without significant channel or background noise effects. Spoofed speech is modified from the original genuine speech data by using a number of speech synthesis (SS) and voice conversion (VC) algorithms. More details and protocols to generate the spoofed speech can be found in. The full dataset is partitioned into three subsets, the first set for training, the second for development and the third for evaluation. The number of speakers and trials in each subset is illustrated in Table 1. There is no speaker overlap across the three subsets.

3 Challenge results

Participants were able to submit scores for up to six systems. One of these systems was designated as the *primary submission*. Spoofing detectors for all primary submissions were trained using only the training data in the ASVspoof 2015 corpus. The dataset was requested by 28 teams from 16 countries; 16 teams returned primary submissions by the deadline. A total of 27 additional submissions were also received. Anonymous results were subsequently returned to each team who were then invited to submit their work to the ASVspoof special session for INTERSPEECH 2015.

Table 1: Number of non-overlapping target speakers and utterances in the training, development and evaluation sets. The duration of each utterance is in the order of one to two seconds.

Subset	#Speakers		#Utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

Table 2: Summary of primary submission results in the ASVspoof 2015 challenge.

System ID	Equal Error Rates (EERs)		
	Known attacks	Unknown attacks	Average
A	0.408	**2.013**	**1.211**
B	0.008	3.922	1.965
C	0.058	4.998	2.528
D	**0.003**	5.231	2.617
E	0.041	5.347	2.694
F	0.358	6.078	3.218
G	0.405	6.247	3.326
H	0.670	6.041	3.355
I	0.005	7.447	3.726
J	0.025	8.168	4.097
K	0.210	8.883	4.547
L	0.412	13.026	6.719
M	8.528	20.253	14.391
N	7.874	21.262	14.568
O	17.723	19.929	18.826
P	21.206	21.831	21.518
Average	3.337 (STD: 6.782)	9.294 (STD: 6.861)	6.316 (STD: 6.558)

This paper summarises the challenge results for primary submissions only. EER results are illustrated in Table 2 in which each line represents the submission of each team. Results are shown independently for known attacks (S1-S5), unknown attacks (S6-S10) and the average, ranked according to the latter. Almost all submissions achieved excellent performance for known attacks (for which training data was provided). EERs in the case of unknown attacks are significantly and universally higher. The lowest EER for all attacks is 1.211%, whereas those for known and unknown attacks are 0.003% and 2.013%, respectively. The lowest EER for unknown attacks (2.013%) is 671 times higher than that for known attacks (0.003%).

These results illustrate the potential of over-fitting countermeasures to known attacks which may leave ASV systems prone to unforeseen spoofing attacks. For example, while system **D** achieves a lower EER than system **A** in the case of known attacks (0.003% vs 0.408%), the EER for system **D** is over twice that of system **A** in the case of unknown attacks (5.231% vs 2.013%). In turn these results thus confirm the importance of developing more generalised countermeasures and also the need for further work and future evaluations.

4 Conclusions

The first automatic speaker verification spoofing and countermeasures challenge (ASVspoof 2015) was highly successful in attracting significant participation. This paper presents the challenge database, organisation, evaluation results, and priorities for future challenges and research.

5 References

[1] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66(0):130 – 153, 2015.

[2] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proceedings of INTERSPEECH*, 2015.

[1]This is a significantly short version of [2] for the UK speech.
[2]http://www.spoofingchallenge.org

Audio Features for Classification of Engagement

Christy Elias, João P. Cabral, Nick Campbell

School of Computer Science and Statistics, Trinity College Dublin

eliasc@tcd.ie, cabralj@tcd.ie, nick@tcd.ie

Abstract

Engagement between interlocutors in a conversation can be estimated from features encapsulated in the speech signal. The features including prosodic (F0), glottal prameters correlated with voice quality (OQ, RQ, SQ), and Mel-frequency cepstral coefficients (MFCCs) are used as parameters to classify engagement in a multi-party dialogue corpus (TableTalk corpus). Combination of these features were used in a random forest classifier and results show that the use of voice quality features improves the classification results.

Index Terms: engagement classification, voice quality

1. Introduction

Engagement detection can be applied to improve the quality of interactions in dialogue systems or to make human intervention decisions in automated telephone calls, etc. Researchers have used different methods to detect engagement in the past, Yu et. al proposed a method for detecting user engagement that uses a multilevel structure with acoustic, temporal and emotional information [1]. Gustafson and Neiberg used prosodic cues from the non lexical response tokens in swedish to detect engagement [2]. Gatica-Perez used the term interest to designate people's internal states related to the degree of engagement displayed [3] and used multimodal cues to detect engagement in multiparty meetings. Bohus and Horvitz modelled engagement in dynamic environments where the participants enter, leave and interact in a very natural manner [4]. Although engagement has been studied previously the features used as cues or the context of the conversations is different in each.

Engagement is a complex concept and in the context of this study an interlocutor is considered to be engaged if he/she is in overall involved in the conversation, interacting with others and/or actively listening to others. The data used in the study is part of the TableTalk[1] corpus , which was collected at ATR Research Labs in Japan for studies in social conversations. Bonin et al. annotated the day 1 recording (35 minutes long) for individual and group engagement with discrete markings of "+" for engaged and "+" for not engaged, the annotators were given no restriction on the length of each annotated segments and the group was considered to be engaged if three out of the four participants in the conversations were engaged [5].

2. Engagement Classification

The speech features were extracted on frames 25 ms long and using a frame shift of 5ms. F0 and MFCC correlates were extracted using the SPTK toolkit (http://sp-tk.sourceforge.net). Voice quality features (glottal parameters), open quotient (OQ), return quotient (RQ) and speech quotient (SQ), were estimated using the method described in [6].

[1] http://sspnet.eu/2010/02/freetalk/

New labels were derived from the original annotations by combining the labels of the different annotators. Each segment was given a label engaged/not-engaged based on the maximum number of annotations that were marked as engaged/not-engaged (respectively) in overlapping segments.

A random forest classifier with 10-fold cross-validation was used to evaluate how the different combinations performed. In order to account for the imbalance of engaged and not-engaged instances in the feature set, weighted accuracy was calculated for each of the classifications. The combination of MFCC, F0 and Voice Quality features resulted in the higher accuracy among the combinations used (88.24%). The results were statistically significant with p-value$<$0.05

3. Acknowledgements

This research is supported by the Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

4. References

[1] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *Computing Research Repository*, vol. cs.SD/0410027, 2004.

[2] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in swedish," in *Proceedings of DiSS-LPSS Joint Workshop 2010*, 2010, pp. 63–66.

[3] D. Gatica-Perez, "Modeling interest in face-to-face conversations from multimodal nonverbal behavior," in *Multimodal Signal Processing*. Academic Press, 2009, pp. 309–326.

[4] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2009, pp. 225–234.

[5] F. Bonin, R. Bock, and N. Campbell, "How do we react to context? annotation of individual and group engagement in a video corpus," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, Sept 2012, pp. 899–903.

[6] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Proc. of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany*, 2007, pp. 113–118.

Towards giving people with physical impairments a voice entry to the digital world

Bahman Mirheidari, Heidi Christensen, Phil Green

Computer Science, University of Sheffield, United Kingdom
{bmirheidari1, heidi.christensen, phil.green}@sheffield.ac.uk

Abstraction

Computers nowadays have influenced almost all aspects of our lives ranging from education and entertainment to communication, social lives and work. However, there are still a significant number of people who cannot access the digital world due to their physical conditions. Speech recognition technology can potentially be used as an option to help people with disabilities to access computers, however, there are people with speech disorders, in particular people with *dysarthria*, who cannot use the current, mainstream speech recognition technology.

The homeService project, developed at University of Sheffield aims to apply state-of-the-art speech technology to help people with dysarthria to enjoy a better life. The system allows people to use speech recognition in their homes to control devices in their environment ("environmental control") such as TV, lights, phone, etc. The long-term goal of the current work is to ultimately use the homeService speech recognition technology to help people with dysarthria have access to computers and use speech recognition as an alternative to the conventional keyboard-mouse method.

Two alternatives tools are designed and implemented: V-Mouse (voice controlled mouse) and V-Keyboard (Voice controlled keyboard). V-Mouse allows clients to move the mouse cursor over the computer screen and click on a desired item using commands like up, down, left, right, and click. Anther approach is to split the screen into smaller areas, e.g. a 3 by 3 grid and move the mouse cursor to an area. Each area can be divided, in turn, into further inner grids (second level grid, third level and so on). V-Keyboard displays a virtual keyboard on the screen and lets the client hit the keys and type words or phrases on text boxes, files or editors. The preliminary tests show significant advantages of the system over the commonly used scanning techniques in terms of time taken to complete simple PC tasks such as navigating to a particular web-page, and inputting a search string.

Automatic speech recognition for people with disordered speech: results from online and offline experiments

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

[1]Computer Science; University of Sheffield, United Kingdom
[2]Human Communication Sciences, University of Sheffield, United Kingdom

The homeService research project is concerned with developing personalised speech-enabled interfaces for users with severe physical impairments and associated disordered speech. By putting state-of-the-art speech recognition systems into people's homes, during long-term trials, invaluable lessons can be learned from doing research 'in-the-wild'.

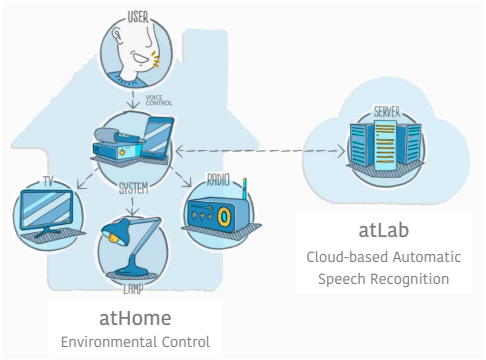


Figure 1: Diagram of the homeService system with its two distinct parts: the atHome component in a user's home and the at- Lab 'in-the-cloud' part. Even though only one user is drawn, the cloud-based ASR server enables simultaneous speech recognition from many users.

Each homeService system is initially deployed with acoustic models adapted using a relatively small amount of enrolment data. During use, data is subsequently collected as the user interacts with the system and this data is used to update the models at a later stage.
This paper contrasts results from experiments carried out online, with the live system, and offline with the collected data. Particular emphasis is put on the amount of adaptation data as well as the use of manual vs. automatic annotations in the context of trying to ensure that the implementation and personalisation strategy will scale with many users.
The audio data that was used in this paper consists of audio files recorded in real home environment by a single user, *M02*. He has motor neuron disease and moderate speech impairment. His system is set up to enable him to control his TV and skybox with speech commands.
Results of the experiments presented in this paper are displayed in Figure 2.

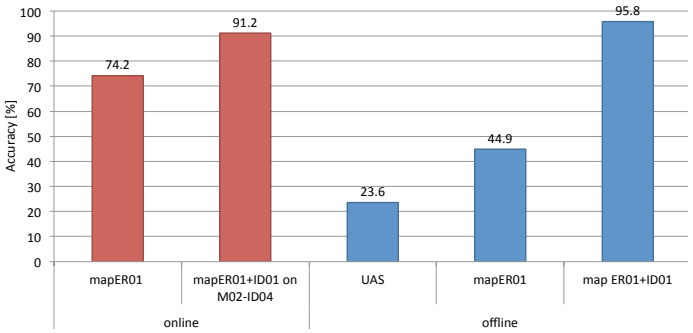


Figure 2: Comparison among different acoustic model in online and offline experiments on M02-ID02 dataset except where stated. Please note that the online test with newly adapted model could not be tested on the same data, due to online tests characteristics.

From the figure, it is clear how acoustic model adaptation is fundamental to improve system performance. Offline experiment accuracy is boosted from 23.6%, corresponding to a generic dysarthric speech model (UAS), to 44.9% with adaptation on small amount of data (3 minutes of user's speech, *ER01*), and to 95.8% with adaptation on a larger recorded dataset (56 minutes, *ID01*).
Same performance improvement is observed in the online experiments, in which accuracy increases from 74.2% to 91.2% by using the best model tuned in the offline experiments

Automatic Dialect Detection and Identification of Code-Switching in Arabic Broadcast Speech

Ahmed Ali, Peter Bell, Steve Renals
Centre for Speech Technology Research, School of Informatics
University of Edinburgh, Edinburgh EH8 9AB, UK
{ahmed.ali, peter.bell, s.renals}@ed.ac.uk

Abstract: We investigate different approaches for dialect identification and code-switching in Arabic broadcast speech, using phoneme sequence, phoneme duration and lexical features obtained from a speech recognition system. We combined these features using a multi-class Support Vector Machine classifier. We validated our results on an Arabic/English language identification task, with an accuracy of 98%. We used these features in a binary classifier to discriminate between Modern Standard Arabic (MSA) and Dialectal Arabic (DA), with an accuracy of 96%. We further report results using the proposed method to discriminate between the five most widely used dialects of Arabic: namely Egyptian, Gulf, Levantine, North African, and MSA, with an overall precision of 49%. We discuss dialect identification error in the context of dialect code- switching between DA and MSA, and compare the error pattern between manually labeled data, and the output from our classifier.

	EGY	GLF	LAV	MSA	NOR	#Truth	%Precision
EGY	**157**	51	52	16	39	315	**49.56%**
GLF	42	**121**	46	29	27	265	**35.07%**
LAV	62	78	**127**	37	44	348	**42.19%**
MSA	14	20	23	**211**	15	283	**63.75%**
NOR	41	75	53	38	**148**	355	**54.21%**
#Classified	316	345	301	331	273		
%Recall	**49.84%**	**45.66%**	**36.4%**	**74.56%**	**41.69%**		

Table 1: Confusion matrix for dialect recognition

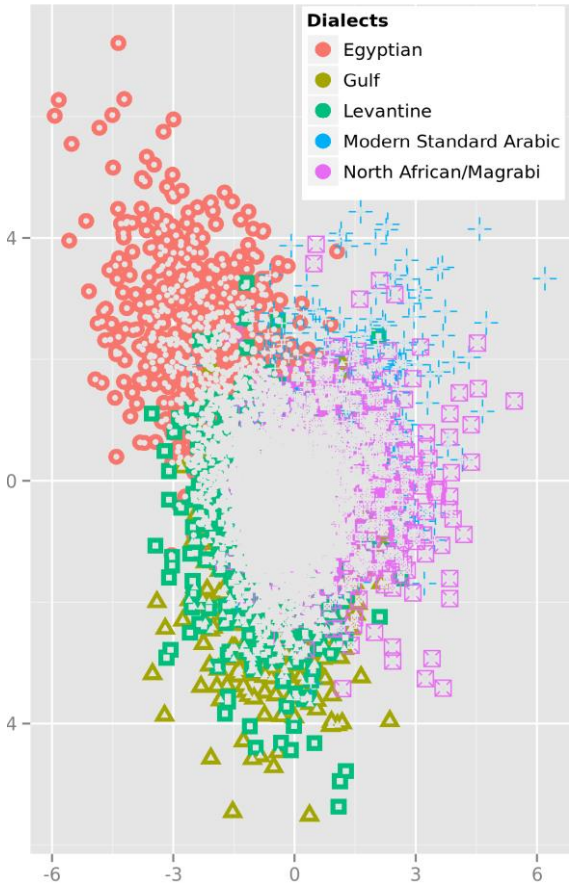


Figure 1: LDA projection for Dialectal Features.

Modelling the effects of hearing aid algorithms on hearing impaired-listeners' perception of speaker intelligibility in noise

Lindon Falconer[1,2], Andre Coy[2], Jon Barker[1]

[1]University of Sheffield, UK; [2]University of the West Indies, Jamaica

[1]{lwfalconer1, j.p.barker}@sheffield.ac.uk; [2]{lindon.falconer, andre.coy02}@uwimona.edu.jm

This study investigates the tuning of hearing aid signal processing algorithms, by using a computational model which predicts the intelligibility of different speakers in noise as perceived by sensorineural hearing impaired (SHI) listeners. The tuning involves determining parameters of the algorithms that enhance speakers' intelligibility as perceived by a particular SHI listener. The computational model consists of two stages. First, an auditory-based model of hearing impairment produces a spectral-temporal excitation pattern (STEP) and time-frequency mask that mimic the measured hearing threshold shifts and loudness recruitment of a specific SHI individual. Second, a microscopic intelligibility model that uses statistical speech models and knowledge of the background noise makes specific predictions about the words that the listener will hear given the STEP. Preliminary experimental results using the Grid corpus show that the model predicts an increased in speech intelligibility for SHI listeners using the NAL-RP prescription. A linear frequency lowering algorithm shows small increase in speaker intelligibility for hearing impairment with steeply sloping hearing loss. This increase occurs for the recognition of utterances of some letters containing high frequency fricatives and unvoiced stops.

Using Sub-Word Hidden Markov Models for Speech Enhancement

Akihiro Kato and Ben Milner

University of East Anglia
Akihiro.Kato@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work proposes a method of speech enhancement based on using a network of HMMs to synthesise clean speech from a noisy input signal. Input speech is decoded by the network of HMMs to provide a model and state sequence. Different choices of acoustic model are considered (whole-word, monophone and triphone) and different grammars (highly constrained to unconstrained). The model and state sequence is applied to the HMMs to synthesise a clean time-frequency surface that is input into a speech production model to produce the enhanced speech signal. Fundamental frequency and voicing can also be synthesised by the HMMs or estimated from the noisy speech. A PESQ analysis compares performance using whole-word, monophone and triphone HMMs along with constrained and unconstrained grammars, and also includes conventional enhancement methods for comparison.

Index Terms: speech enhancement, HMMs, STRAIGHT

Evidence of Phonological Processes in Automatic Recognition of Children's Speech

Eva Fringi, Jill Fain Lehman, Martin Russell

Abstract

Children's speech is characterised by high acoustic and linguistic variability [1, 2, 3]. This makes automatic speech recognition (ASR) for children's speech more difficult than for adults' speech [4, 5, 6]. A plausible explanation is that ASR errors are due to predictable phonological effects associated with language acquisition [7, 8, 9, 10, 11]. We describe phone recognition experiments on hand labelled data for children aged between 5 and 9. A comparison of the resulting confusion matrices with those for adult speech (TIMIT [12]) is conducted with the use of a statistical significance test and shows increased phone substitution rates for children, which correspond to some extent to established phonological phenomena. However these errors still only account for a relatively small proportion of the issue. This suggests, in agreement with previous studies [13, 14], that attempts to improve ASR accuracy on children's speech by accommodating these phenomena, for example by changing the pronunciation dictionary, cannot solve the whole problem. Further investigation is proposed with the use of a more robust ASR system which would potentially decrease the general substitution rate and make the language acquisition related errors more prominent.

[1] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," Journal of the Acoustical Society of America, vol. 105, no. 3, pp. 1455–1468, 1999.
[2] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in Proc. IEEE-ICASSP, Honolulu, Hawai, 2007.
[3] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analysing children's speech: An acoustic study of consonants and consonant-vowel transition," in Proc. IEEE-ICASSP, Toulouse, France, vol. 1, 2006.
[4] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in Proc. IEEE-ICASSP, Atlanta, GA, 1996.
[5] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in FONETIK 2004, 2004.
[6] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in Proc. IEEE-ICASSP, Hong Kong, 2003.
[7] B. Lust, Child Language: Acquisition and Growth. University Press, 2006. Cambridge
[8] B. Smit, A., J. J. Hand, L. andFreilinger, E. Bernthal, J., and A. Bird, "The iowa articulation norms project and its Nebraska replication," vol. 55, 1990.
[9] B. Dodd, A. Holm, Z. Hua, and S. Crossbie, "Phonological development:a normative study of british-english speaking children," Clinical Linguistics and Phonetics, vol. 17, no. 8, pp. 617–643, 2003.
[10] W. Cohen and C. Anderson, "Identification of phonological processes in preschool children's sinlge-word productions," International Journal of Language and Communication Disorder, vol. 46, no. 4, pp. 481–488, 2011.
[11] S. McLeod and J. Arciuli, "School-aged children's production of /s/ and /r/ consonant clusters," vol. 61, pp. 336–341, 2009.
[12] J. S. Garofolo et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.
[13] P. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children's speech using acoustic adaptation and pronunciation modeling," in Proc. Workshop on Child-Computer Interaction, WOCCI, 2014.
[14] Q. Li and M. Russell, "An analysis of the causes of increased error rates in childrens speech recognition," in Proc. ICSLP, 2002.

Automatic speech recognition and keyword search for low-resource languages: Babel project research at CUED

Mark Gales, Kate Knill, Anton Ragni, Haipeng Wang, Phil Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK

Abstract

Recently there has been increased interest in Automatic Speech Recognition (ASR) and Key Word Spotting (KWS) systems for low resource languages. One of the driving forces for this research direction is the IARPA Babel project. This paper describes some of the research funded by this project at Cambridge University, as part of the Lorelei team co-ordinated by IBM. A range of topics are discussed including: graphemic lexica; sub-word language models; efficient decoding using multiple deep neural network based acoustic models; and zero acoustic model resource systems. Performance for all approaches is evaluated using the Very Limited (approximately 3 hours) and/or Full (approximately 80 hours) language packs distributed by IARPA. Both KWS and ASR performance figures are given. Though absolute performance varies from language to language, and keyword list, the approaches described show consistent trends over the languages investigated to date. Using comparable systems over the six Option Period 2 languages indicates a strong correlation between ASR performance and KWS performance.

Reconstructing Voices within the Multiple-Average-Voice-Model Framework

Pierre Lanchantin[†], Christophe Veaux[*], Mark J.F. Gales[†], Simon King[*], Junichi Yamagishi[*]

[†]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

{pkl27,mjfg}@cam.ac.uk

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

cveaux@inf.ed.ac.uk, simon.king@ed.ac.uk, jyamagis@inf.ed.ac.uk

Degenerative speech disorders can be due to a variety of causes including Multiple Sclerosis, Parkinson's and Motor Neurone Disease (MND). Initial symptoms of MND may include reduction in speaking rate, increase of voice's hoarseness and/or imprecise articulation. As the disease progresses, most patients become unable to meet their daily communication needs using speech and most are unable to speak by the time of their death. As speech becomes unintelligible, voice output communication aids (VOCAs) may be used. These devices consist of a text entry interface (keyboard, eye-tracker) and a text-to-speech synthesizer that generates the corresponding speech. VOCAs are usually limited to a set of impersonal voices that not match necessarily the individual in terms of age or accent, which can cause embarrassment and a lack of motivation to interact socially [1]. In fact, speech synthesis is not just an optional extra for reading out text, but a critical function for social communication and personal identity. Hence, provision of personalised voice is associated with greater dignity and improved self-identity for the individual and their family [1].

Most existing personalised VOCA devices (ModelTalker[2], Cereproc[3], Polluxstar, based on a hybrid TTS [4] using both unit selection and statistical parametric speech synthesis [5]) are based on a voice banking approach which is the process of capturing the voice before it starts to degrade. They require a large amount or recorded intelligible speech (before degradation) in order to build a good quality voice. This is problematic for patients whose voices have already started to deteriorate and there is a strong motivation to reduce complexity and to increase the flexibility of the voice building process so that patients can have their own synthetic voices built from limited recordings and even deteriorating speech. HMM-based speech synthesis techniques have recently been used to create personalised VOCAs [6, 7]. One advantage is speaker adaptation [8] of pre-trained Average Voice Model (AVM) towards a target speaker which allows the construction of voices from limited recordings. An other advantage is linked to the statistical nature of the approach which allows voice reconstruction ([9, 10]) via the control/modification of various components to compensate for the disorders found in the patient's speech. Hence, considering statistically independent models for duration, log-f_0, band aperiodicity and mel-cepstrum, a possible approach proposed in [6] involves the substitution of some models in the patient's speaker-adapted voice by that of a well-matched healthy voice. Knowing that articulatory errors in disordered speech are consistent [11] and hence relatively predictable [12], substitution strategy can be pre-defined for a given condition. For instance, speaking rate is a common disorder of MND patient's speech which can lead to a loss of speech intelligibility. Substituting the state duration models enables the timing disruptions to be regulated. Breathy or hoarse speech is an other common disorder. In such cases, a possible strategy is to substitute the band aperiodicity models. However, each substitution might remove some of the identity of the speech and it is crucial to preserve components which are highly correlated with the speaker identity.

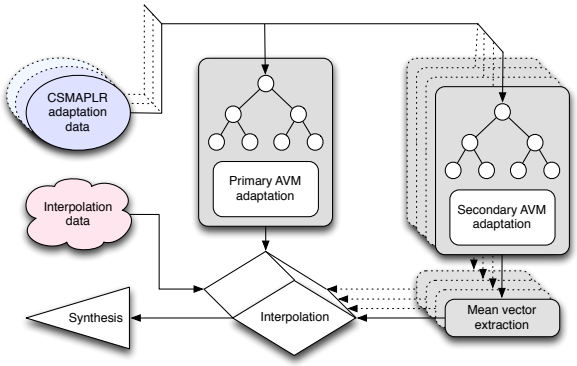


Figure 1: The Multiple-AVM framework.

The Multiple-AVM approach was recently introduced in [13]. It can be seen as an hybrid between the AVM [8] and the Cluster Adaptive Training (CAT [14]) approaches. The adaptation procedure is illustrated in Figure 1. In the same fashion than CAT, during the adaptation of a Gaussian component, the set of adapted AVM mean vectors constitutes an "eigenspace"[1] in which the adapted mean vector of the component is interpolated. However, clusters are AVMs which can be adapted so that the eigenspace can be tuned towards the target voice before interpolation. As in the (single-) AVM approach, each AVM is pre-trained independently on a selection of speakers from a voice bank and decision trees of the considered AVMs can be intersected during interpolation, allowing a wider variety of possible contexts to be produced. In this paper we show that this framework is well-suited to the voice reconstruction task, both in terms of complexity and flexibility of the creation process. For instance, the eigenspace can be designed using different combinations of AVMs/target voices and the interpolation can be done in a "clean" space [15] by selecting healthy target voices close to the disordered one. Moreover, the interpolation weights distribution can be fine-tuned manually after interpolation by a practician, according to the speaker's or to his family's appreciation. Finally the interpolation can be performed with only a small amount of adaptation data as it only requires the estimation of the weights interpolation vector. We illustrate our points with a subjective assessment of the reconstructed voice.

This research was supported by ESPRC Programme Grant, grant no. EP/I031022/1 (Natural Speech Technology)

[1] no orthogonality constraints are considered here.

1. References

[1] J. Murphy, "I prefer this close': Perceptions of AAC by people with motor neurone disease and their communication partners," *Augmentative and Alternative Communication*, vol. 20, pp. 259–271, 2004.

[2] D. Yarrington, C. Pennington, J. Gray, and H. Bunnell, "A system for creating personaized synthetic voices," in *Proc. of ASSETS*, 2005.

[3] http://cereproc.com.

[4] H. Kawai, K. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and K. Tokuda, "XIMERA: a concatenative speech synthesis system with large-scale corpora," *IEICE Trans. Information and Systems*, no. J89-D-II(12), pp. 2688–2698, 2006.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] S. Creer, P. Green, S. Cunningham, and J. Yamagishi, "Building personalized synthesized voices for individuals with dysarthia using the HTS toolkit," in *IGI Global Press*, Jan. 2010.

[7] Z. Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the voice of an individual following laryngectomy," in *Augmentative and Alternative Communication*, 2011.

[8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, January 2009.

[9] C. Veaux, J. Yamagishi, and S. King, "Voice banking and voice reconstruction for MND patients," in *Proc. of ASSETS*, 2011.

[10] ——, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *Proc. Interspeech*, 2012.

[11] K. M. Yorkston, D. R. Beukelman, and K. R. Bell, "Clinical management of dysarthric speakers," in *College-Hill Press*, 1998.

[12] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. ICASSP 2011*, 2011.

[13] P. Lanchantin, M. J. F. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," in *Proc. ICASSP*, 2014.

[14] M. Gales, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 8, pp. 417–428, 2000.

[15] K. Yanagisawa, J. Latorre, V. Wan, M. J. F. Gales, and S. King, "Noise robustness in HMM-TTS speaker adaptation," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

Towards minimum perceptual error training for DNN-based speech synthesis

Cassia Valentini-Botinhao, Zhizheng Wu, Simon King

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

cvbotinh@inf.ed.ac.uk {zhizheng.wu, simon.king}@ed.ac.uk

Abstract

The quality of speech generated by statistical parametric systems has benefited from advances in acoustic models, vocoders and postfilters. However the challenge of how to create truly high quality speech from learned vocoder parameters still remains. The vocoder itself is certainly one of the main limitations. But modelling assumptions, such as independence among different acoustic parameters, e.g., source and the filter, have also been shown to cause great degradation [1]. It is inevitable that any vocoder or statistic model will introduce error, so perhaps we should aim for errors that are introduced at any stage of the process to be as imperceptible as possible.

The idea of minimising a perceptual error is not new. Minimum Generation Error (MGE) [2, 3] for hidden Markov model (HMM)-based speech synthesis could be thought of as a step in this direction. Unified feature extraction and model training could also lend itself to perceptual error minimisation [4, 5]. Nakamura et. al [4] proposed to extract Mel cepstral coefficients that maximize the likelihood of the data. More recently Shinji et. al introduced a compact representation of the spectrum using autoencoders [5]. Both techniques could be seen as error-minimising alternatives to Mel cepstral analysis [6].

The recent success of Deep Neural Network (DNN) speech synthesis [7–10] suggest a range of new directions for minimum perceptual error training. In general, when training a DNN to predict acoustic parameters, all parameters are optimised using a shared cost function, allowing the model potentially to learn dependencies between output parameters.

DNN training easily allows for different cost functions to be used. It is possible to train a DNN to predict Mel cepstral coefficients but to calculate the error in the higher-dimensional spectral domain, simply by reformulating the cost function. It is also possible to train a DNN to predict the spectrum directly.

There are, however, more perceptually relevant representations of speech that could be used to measure the error, but that do not allow for synthesis. So, we might measure the error not directly on the output acoustic features (i.e., vocoder parameters) but in some other domain, which may not itself be useful for speech generation. In this situation, it is desirable to train a model that predicts vocoder parameters – necessary to eventually generate the waveform – but to calculate the error in this perceptual domain. In this paper we exploit this idea, using a particular perceptual representation of the speech spectrum.

We propose to use a perceptually-oriented domain to improve the quality of text-to-speech generated by deep neural networks (DNNs). We train a DNN that predicts the parameters required for speech reconstruction but whose cost function is calculated in another domain. In this paper, to represent this perceptual domain we extract an approximated version of the Spectro-Temporal Excitation Pattern that was originally proposed as part of a model of hearing speech in noise. We trained DNNs that predict band aperiodicity, fundamental frequency and Mel cepstral coefficients and compare generated speech when the spectral cost function is defined in the Mel cepstral, warped log spectrum or perceptual domains. Objective results indicated that the perceptual domain system achieves the highest quality. Calculating the cost in the spectrum domain however generated speech that was most preferred by listeners. Future work includes considering different perceptual domains.

1. References

[1] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, vol. 15, September 2014, pp. 1504–1508.

[2] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-Based speech synthesis," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. I89–I92.

[3] Y.-J. Wu and K. Tokuda, "Minimum generation error training by using original spectrum as reference for log spectral distortion measure," in *Proc. ICASSP*, Taipei, Taiwan, April 2009, pp. 4013–4016.

[4] K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Integration of acoustic modeling and mel-cepstral analysis for hmm-based speech synthesis," in *Proc. ICASSP*, May 2013, pp. 7883–7887.

[5] S. Takaki and J. Yamagishi, "Constructing a deep neural network based spectral model for statistical speech synthesis," in *NOLISP (submitted)*, 2015.

[6] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, San Francisco, USA, March 1992, pp. 137–140.

[7] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7962–7966.

[8] Y. Qian, Y. Fan, W. Hu, and F. Soong, "On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis," in *Proc. ICASSP*, May 2014, pp. 3829–3833.

[9] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florance, Italy, 2014, pp. 3872–3876.

[10] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, 2015.

Acknowledgements This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

I-Vector Estimation Using Informative Priors for Adaptation of Deep Neural Networks

Penny Karanasou, Mark Gales, Philip Woodland

22 June 2014

I-vectors are a well-known low-dimensional representation of speaker space and are becoming increasingly popular in adaptation of state-of-the-art deep neural network (DNN) acoustic models. One advantage of i-vectors is that they can be used with very little data, for example a single utterance. However, to improve robustness of the i-vector estimates with limited data, a prior is often used. Traditionally, a standard normal prior is applied to i-vectors, which is nevertheless not well suited to the increased variability of short utterances. This is because of a Gaussian assumption over the i-vector space which is not always true; when reducing duration, the variance of the i-vector estimate increases and decisions become error-prone.

This paper proposes a more informative prior, derived from the training data. As well as aiming to reduce the non-Gaussian behaviour of the i-vector space, it allows prior information at different levels, for example gender, to be used. A count-smoothing framework is adopted for incorporating the prior knowledge into i-vector estimation. The smoothing idea is based on the interpolation of observed statistics and prior statistics, both derived from the training data. In this work, the prior statistics were first estimated across all training speakers offering an average representation of the speaker space. Second, gender clustering of the training data was used and the prior statistics of two gender i-vectors were extracted. Our approach integrates prior estimation into EM training of the i-vectors after a normalisation of the prior statistics. A normalised prior estimated using the training data models the actual behaviour of the speaker space and is less sensitive to the quantity used to estimate the i-vector and to the mismatch between training and test data.

Experiments on a US English Broadcast News (BN) transcription task for speaker and utterance i-vector adaptation show that more informative priors reduce the sensitivity to the quantity of data used to estimate the i-vector. The best configuration for this task was utterance-level test i-vectors enhanced with informative priors which gave a 13% relative reduction in word error rate over the baseline (no i-vectors), and a 5% relative reduction in word error rate over utterance-level test i-vectors with standard prior.

Paraphrastic Recurrent Neural Network Language Models,
Xunying Liu, Xie Chen, Mark Gales and Phil Woodland.

Abstract Recurrent neural network language models (RNNLM) have become an increasingly popular choice for state-of-the-art speech recognition systems. Linguistic factors influencing the realization of surface word sequences, for example, expressive richness, are only implicitly learned by RNNLMs. Observed sentences and their associated alternative paraphrases representing the same meaning are not explicitly related during training. In order to improve context coverage and generalization, paraphrastic RNNLMs are investigated in this paper. Multiple paraphrase variants were automatically generated and used in paraphrastic RNNLM training. Using a paraphrastic multi-level RNNLM modelling both word and phrase sequences, significant error rate reductions of 0.6% absolute and perplexity reduction of 10% relative were obtained over the baseline RNNLM on a large vocabulary conversational telephone speech recognition system trained on 2000 hours of audio and 545 million words of texts. The overall improvement over the baseline n-gram LM was increased from 8.4% to 11.6% relative.

Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics

Linxue Bai, Peter Jančovič, Martin Russell, Philip Weber

School of EESE, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

{lxb190,p.jancovic,m.j.russell}@bham.ac.uk,phil.weber@bcs.org.uk

Keywords: modelling speech dynamics, neural networks, bottleneck features, CSHMM, low-dimensional features, formants.

1. Introduction

Segmental models of speech (e.g. [1]) hold promise for speech recognition due to their ability to parsimoniously model speech dynamics. However they have been hampered by lack of a good feature representation. MFCCs perform well for speech recognition, but are less suitable for segmental models, since the articulator dynamics of speech are manifested indirectly, often as movement between, rather than within, frequency bands. Formants model voiced sounds well, but are difficult to estimate reliably and are inappropriate for unvoiced speech, while articulatory parameters are difficult to obtain. Therefore, there is a need for a compact representation of speech, suitable for segmental models, that can be estimated for all speech sounds.

Recent research has used (deep) neural networks (NNs) with a 'bottleneck' compression layer as a non-linear feature extractor for speech recognition, typically containing tens to hundreds of neurons (e.g. [3, 2]). Low-dimensional representations instead reflect that the mechanisms of speech production involve movement of a small number of speech articulators.

We present an analysis of a low-dimensional representation of speech for modelling speech dynamics, extracted from a bottleneck layer with a small number (3 to 12) of neurons. The input to the network is a set of spectral feature vectors. We explore the effect of various designs and training of the network: varying the size of context in the input layer, size of the bottleneck and other hidden layers, and training the network to reconstruct the input, or to predict phone posterior probabilities.

We employ our bottleneck features in a conventional HMM-based phoneme recognition system, achieving recognition accuracy of 70.6% on the TIMIT core test set using only 9-dimensional features (Table 1). We show an average 33.7% reduction in phone errors compared with employing formant-based features of the same dimensionality. These experiments used features obtained from the middle layer of a network having five layers (layers 2 and 4 contained 512 neurons), trained on log filter-bank energies with 5 frames preceding and succeeding context (286 dimensions), to predict posterior probabilities of 49 phoneme classes. Scoring used 40 classes.

We also analyse how the bottleneck features fit the assumptions of dynamic models of speech. Specifically, we employ the continuous-state Hidden Markov Model (CSHMM) [4], which considers speech as a sequence of alternating dwell and transition regions (proposed by Holmes, Mattingley and Shearme [5]) modelled by smoothly-varying features such as formants. A sequential branching algorithm recovers the sequence of dwells and transitions, times of changes between them, and the sequence of phonemes which could have generated them.

Feature representation	Dim.	Corr (%)	Acc (%)
Baseline: MFCC + Δ + ΔΔ	39	76.2	71.0
3 formants	3	49.3	40.7
3 formants & amp & bw+ Δ + ΔΔ	27	65.1	60.4
3 BN features	3	65.0	60.9
9 BN features	9	74.4	70.6
9 BN features + Δ + ΔΔ	27	76.8	73.1

Table 1: *Recognition performance of an HMM-based ASR system when using formant and bottleneck features extracted from a network trained to predict phoneme posterior probabilities.*

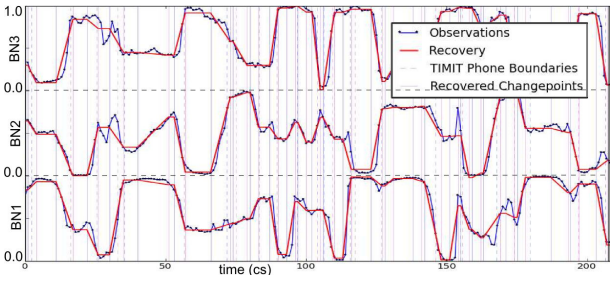


Figure 1: *Example dwell-transition trajectories (red) recovered by the CSHMM using bottleneck features (blue). Vertical lines show TIMIT phone boundaries. and recovered dwells.*

2. Conclusion

We show that bottleneck features, from networks trained to predict phoneme posteriors, significantly outperform formant features of similar dimensionality. They also preserve the trajectory continuity well (better than formants), thus providing a suitable representation for decoding with the CSHMM.

3. References

[1] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, 4(5), pp. 360–378, 1996.

[2] L. Deng and J. Chen, "Sequence classification using the high-level features extracted from deep neural networks," in *proc. ICASSP*, pp. 6844–6848, 2014.

[3] D. Liu, S. Wei, W. Guo, Y. Bao, S. Xiong, and L. Dai, "Lattice based optimization of bottleneck feature extractor with linear transformation," in *proc. ICASSP*, pp. 5617–5621, 2014.

[4] C. J. Champion and S. M. Houghton, "Application of Continuous State Hidden Markov Models to a classical problem in speech recognition," accepted to *CSL*, 2015.

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

Finding phonemes: improving machine lip-reading

Helen L. Bear, Richard W. Harvey, Yuxuan Lan
{helen.bear,r.w.harvey,y.lan}@uea.ac.uk
School of Computing Sciences, University of East Anglia,
Norwich, NR4 7TJ, UK

In machine lip-reading there is continued debate and research around the correct classes to be used for recognition. Here we use a structured approach for devising speaker-dependent viseme classes, which enables the creation of a set of phoneme-to-viseme maps where each has a different quantity of visemes ranging from 2 to 45. Viseme classes are based upon the mapping of articulated phonemes, which have been confused during phoneme recognition, into viseme groups. Using these maps, with a 1000-word dataset, we show the effect of changing the viseme map size in speaker-dependent machine lip-reading, measured by word recognition correctness and so demonstrate that word recognition with phoneme classifiers is not just possible, but often better than word recognition with viseme classifiers. Furthermore, there are intermediate units between visemes and phonemes which are better still.

Oral Session 2

Annotating large lattices with the exact word error

Rogier C. van Dalen, Mark J. F. Gales
Department of Engineering
University of Cambridge, United Kingdom

June 22, 2015

The acoustic model in modern speech recognisers is trained discriminatively, for example with the minimum Bayes risk. This criterion is hard to compute exactly, so that it is normally approximated by a criterion that uses fixed alignments of lattice arcs. This approximation becomes particularly problematic with new types of acoustic models that require flexible alignments. It would be best to annotate lattices with the risk measure of interest, the exact word error. However, the algorithm for this uses finite-state automaton determinisation, which has exponential complexity and runs out of memory for large lattices. This presentation will introduce a novel method for determinising and minimising finite-state automata incrementally. Since it uses less memory, it can be applied to larger lattices.

Consonant Recognition with Continuous-State Hidden Markov Models and Perceptually-Motivated Features

Philip Weber, Colin Champion, Steve Houghton, Peter Jančovič, Martin Russell

School of EESE, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

phil.weber@bcs.org.uk,{c.champion,s.houghton,p.jancovic,m.j.russell}@bham.ac.uk

Keywords: Perceptual Features, Consonants, CSHMM.

1. Introduction

Research into human speech recognition (e.g. [1]) has shown that humans use a relatively small number of basic features of the auditory signal as perceptual cues to perceive and distinguish consonantal sounds. The features most useful for recognition depend on the specific speech sound. Typical ASR features and recognisers however neither vary with the type of sound nor relate directly to perceptual cues.

A key distinguishing characteristic of non-sonorant speech sounds, plosives, fricatives and affricates in particular, is the amplitude and duration of broadband noise in particular frequency bands. This raises the question: if a basic set of features – phoneme duration and mean log energy in a small set of spectral bands – largely enable human discrimination among non-sonorant consonants, could such perceptually-motivated features serve well for machine recognition of these consonants?

We explore this question through classification and recognition experiments on TIMIT, on a limited set of consonants (stops, closures, unvoiced fricatives and affricates). Our motivation is to develop parsimonious models of speech and appropriate modelling algorithms which are faithful to the dynamics of the speech signal and what is known of the mechanisms of the production of different types of speech sound.

Our features for classification and recognition are based on phoneme-specific perceptual cues identified in the literature, e.g. Li and Allen's psychoacoustic experiments on human perception of plosives [2] and fricatives [3]. Log energies are summed over spectral bands given by frequency boundaries interpreted from these cues, averaged over the duration of the phoneme given by TIMIT boundaries, appended with duration.

We compared basic Gaussian classifiers using features of varying dimension extracted using various FFT parameters. The best classification accuracy 72.5% (Table 1) was obtained with 9-dimensional spectral features obtained from FFTs taken over a 2ms window with no overlap or zero padding.

For recognition we use a variant of the Continuous-State HMMs (CSHMMs) proposed in previous work [4] as the appropriate modelling framework for the dynamics of voiced sounds modelled with formant features. A sequence of consonants can be described by a series of 'dwells', sequences of relatively stationary features, with abrupt transitions between phonemes, and decoded using a 'dwell-only' CSHMM.

In Table 1 we show results from decoding with and without a bigram language model, and a comparative result from a single state, single Gaussian, discrete-state HMM with log-normal timing model. The full results suggest that features optimal for human perception also perform best for machine classification, and perform well for recognition. We also relate characteristics of the models learned, for example 'blocking' in covariance matrices, to prior knowledge of human speech perception and production (Figure 1).

Features	%Corr	%Sub	%Del	%Ins	%Err
Gaussian Classifier	72.5	–	–	–	27.5
CSHMM	55.1	24.8	20.1	2.0	46.9
HMM (MFCC)	57.0	27.9	15.1	2.4	45.4
CSHMM with LM	73.1	19.5	8.2	3.2	30.8

Table 1: Phoneme classification and recognition results.

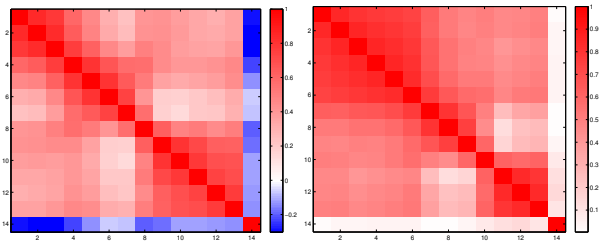


Figure 1: Correlation matrices (spectral and duration features) for /g/ (left panel) and /sh/ (right panel).

2. Conclusion

Our aim is speech recognition based on faithful, parsimonious models of speech. We showed that linguistically meaningful features are suitable for a dwell-only CSHMM to recognise unvoiced segments of speech. This model is appropriate for integration with the dwell-transition CSHMM for sonorant speech [4] and provides a natural framework for including other features of known perceptual importance.

3. References

[1] K. N. Stevens, "Acoustic correlates of some phonetic categories," *JASA*, 68(3), pp. 836–842, 1980.

[2] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *JASA*, 127(4), pp. 2599–2610, 2010.

[3] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," *JASA*, 132(4), pp. 2663–2675, 2012.

[4] C. J. Champion and S. M. Houghton, "Application of Continuous State Hidden Markov Models to a classical problem in speech recognition," Accepted to *CSL*, 2015. doi:10.1016/j.csl.2015.05.001.

Source-filter Separation of Speech Signal in the Phase Domain

Erfan Loweimi, Jon Barker, and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK
{e.loweimi, j.barker, t.hain}@dcs.shef.ac.uk

Deconvolution of the speech excitation (source) and vocal tract (filter) components through log-magnitude spectral processing is well-established and has led to the well-known cepstral features used in a multitude of speech processing tasks. This paper presents a novel source-filter decomposition based on processing in the *phase* domain. The phase spectrum is rarely used in mainstream speech processing. There are three major reasons for its neglect. First, it is generally considered to contain little perceptual information. Second, the phase wrapping phenomenon renders the shape of the phase spectrum chaotic and noise-like. The wrapped spectrum lacks the meaningful trends or extremum points that are helpful when developing a model. Third, it has been shown that the speech phase spectrum is only informative when the speech signal is decomposed into long frames (e.g. 500 ms). This is problematic because using long frames violates the quasi-stationarity assumption that is the key motivation for frame-based speech signal processing.

There are recent studies that provide new evidence for the perceptual importance of the phase spectrum. These studies generally incorporate information from the phase spectrum into an existing magnitude spectrum-based enhancement algorithm and then due to some improvement in the intelligibility/quality of the output signal conclude that phase is of some perceptual significance. So far no model is provided that can show how information is encoded in the phase spectrum. This paper provides such an account in the form of a novel phase-based source-filter model.

We show that separation between source and filter in the log-magnitude spectra is not perfect, leading to partial loss of vocal tract information. It is demonstrated that the same task can be better performed by trend and fluctuation analysis of the phase spectrum of the minimum-phase component of speech, which can be computed via the Hilbert transform. Trend and fluctuation can be separated through low-pass filtering of the phase spectrum, exploiting the additivity of the vocal tract and source responses in the phase domain. This results in separated signals which have a clear relation to the vocal tract and excitation components. The effectiveness of this approach to speech modelling is tested using a speech recognition task. The vocal tract component extracted in this way is employed as the basis of a feature extraction algorithm for speech recognition on the Aurora-2 database. The recognition results show up to 8.5% absolute improvement on average (0-20 dB) in comparison with MFCC features.

Joint Modeling of F0 and Duration in Deep Neural Network Based Speech Synthesis

Srikanth Ronanki, Zhizheng Wu, Robert A. J. Clark
Centre for Speech Technology Research, University of Edinburgh

Fundamental frequency(F0) and duration are two important factors in prosody, and a significant amount of work has been done to model them in statistical parametric speech synthesis. However, conventional techniques assume conditional independence between F0 and duration, and model them separately. This paper proposes an approach to jointly model the high-level behaviour of F0 contours and duration within a deep neural network framework.

1 Relation to prior work

Statistical parametric speech synthesis(SPSS) based on hidden Markov models (HMMs) [4] has flourished for many years now and offers greater flexibility to change its voice characteristics than concatenative speech synthesis approaches. Most recently, neural networks have re-emerged as a potential acoustic model for [3, 5] following their success in deep learning for ASR. However, the naturalness of synthesized speech is still generally neutral in terms of prosody and cannot compete with good unit selection systems. One likely reason for this is that pitch variation continues to be modeled locally at the frame level and these models unable to capture long-term behaviour in the pitch contour. Additionally duration is always predicted first and independently of the pitch.

The Discrete Cosine Transform (DCT) is often used to represent F0 at higher levels, which is able to compactly represent complex contours. The Continuous Wavelet Transform (CWT) has also been proposed to improve F0 within the HMM-framework [2]. Here some improvements were seen in the accuracy of F0 modeling. The work proposed in [1] explored a multi-level representation of F0 by combining both DCT and CWT transforms and modeled at different wavelet scales with each scale representing the variations in F0 contour from utterance/phrase level to phoneme level. However, all these approaches require separate models to predict state-level duration first and then use it here to predict the contour.

The novelty of this work is to jointly model duration and the F0 contour above the frame level using on deep neural networks. For this, the DCT representation of F0 contour, the duration of phone and its states are used as input features to the network. This ensures the training to learn the phoneme duration along with state duration with least deviation in between them. In the current work, we also investigate the use of bottleneck features [3] from the DNN as a richer representation of prosodic context to supplement the input. Since, the proposed approach learns the representation at phoneme level, the addition of bottleneck features provide wider context to help the network learn more accurate longer-term variations in F0 contour.

2 Proposed Joint Modeling of F0 and Duration

Let $x_i = [x_i(1),...,x_i(d_x)]^T$ and $y_i = [y_i(1),...,y_i(d_y)]^T$ be static input and target feature vectors of phoneme i where d_x and d_y denote the dimensions of x_i and y_i, respectively, and T denotes transposition. The input features (x_i) include binary features derived from a subset of the questions used by the decision tree clustering in the HMM system. The target features (y_i) include DCT-parameters of a phoneme F0 contour along with its duration and the time-aligned duration of its five states as shown below:

$$f_{0ERB} = \log_{10}(0.00437 * f_0 + 1) \tag{1}$$

$$c[k] = 2 * w[k] * \sum_{n=0}^{N-1} f_i[n] cos(\frac{\pi(2n+1)(k)}{2N}), \quad 0 \leq k < N \tag{2}$$

where $f_i = [f_0(1),...,f_0(t)]$, t denotes number of frames in phoneme i.

$$y_i = [c[0],...,c[k], p_i, s_i^1,...,s_i^5] \tag{3}$$

where $c[0],...,c[N-1]$ represents the DCT-stylized F0 features, p_i is phoneme duration, $s_i^1,...,s_i^5$ represents duration of states. The DNN is then trained to map the linguistic features of input text to the prosodic features of a speaker, i.e., if $D(x_i)$ denotes the DNN mapping of x_i, then the error of mapping is given by $\varepsilon = \sum \|y_i - D(x_i)\|^2$ is defined as

$$D(x_i) = \widetilde{d}(z_{n+1}) \tag{4}$$

$$z_{n+1} = d(w^{(n)} d(z_n)) \tag{5}$$

$$d(\vartheta) = a\tanh(b\vartheta), \widetilde{d}(\vartheta) = \vartheta \tag{6}$$

where n represents number of hidden layers and $w^{(n)}$ represents the weight matrix of n^{th} hidden layer of the DNN model. We tried different architectures by varying the total number of hidden layers from 3 to 6. The best architecture with minimum generation error was found out to be with 6 layers consisting of 1024 nodes in first two hidden layers and 128 nodes in the subsequent four hidden layers. Since the number of input nodes are more in number compared to output layer, the architecture [1024 1024 128 128 128 128] performed better than all other architectures during training.

In another system, a first DNN with architecture [1024 32 1024 1024 1024 1024] is used to extract 32-dimensional bottleneck features with 10 frames context and are stacked with linguistic features as input to a second DNN with architecture [6*1024] to predict prosody features. The bottleneck features represent activation at the hidden layer for each phoneme. The weights generated during DNN training are used to estimate the contour shape (c_1 - c_8) and $mean f_0$. With the help of IDCT and voiced/Unvoiced (V/UV) values from the baseline DNN, the F0 contour is reconstructed and the output F0 values are normalized to zero based on the predicted value of V/UV (if V/UV < 0.5). A STRAIGHT vocoder is used to synthesize the waveform using the predicted Mel-Cepstral, BAP features from the baseline DNN's frame-by-frame mapping and F0 from the proposed method. A speech database from a British male speaker (Nick) was used in the experiments. Objective evaluation is shown in Table 1.

Table 1: Objective evaluations on Nick Database

	F0 contour(Hz)		Duration(Frames)
Method	**RMSE**	**CORR**	**RMSE**
HMM-GV	9.90	0.782	6.05
Baseline-DNN	9.34	0.812	-
DCT-DNN	9.00	0.818	5.15
DNN-DNN	8.86	0.825	5.24

We have proposed the use of joint modeling of F0 and duration in current state-of-the-art DNN-based speech synthesis. It has been shown that discrete cosine transform is a good representation of F0 contour and can be modeled using deep neural networks much better than frame level F0 modeling. Objective evaluation also suggest that these are quite effective.

3 References

[1] S. Ribeiro Manuel and R. A. J. Clark. A Multi-Level Representation of F0 Using The Continuous Wavelet Transform And The Discrete Cosine Transform. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2015.

[2] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio. Wavelets for intonation modeling in HMM speech synthesis. In *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[3] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[4] T Yoshimura, K Tokuda, T Masuko, T Kobayashi, and T Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *EUROSPEECH*. ISCA, 1999.

[5] H. Zen, A. Senior, and M. Schuster. Statistical Parametric Speech Synthesis Using Deep Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7962–7966, 2013.