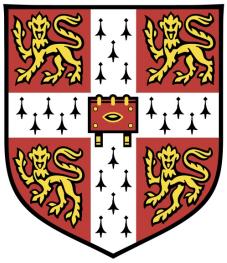
UK Speech Conference

Cambridge

11–12 September 2017

We are grateful for the sponsorship of

amazon



Programme

The technical programme contains 3 keynotes, 2 oral sessions with 5 presentations and 3 poster sessions with 50 total posters.

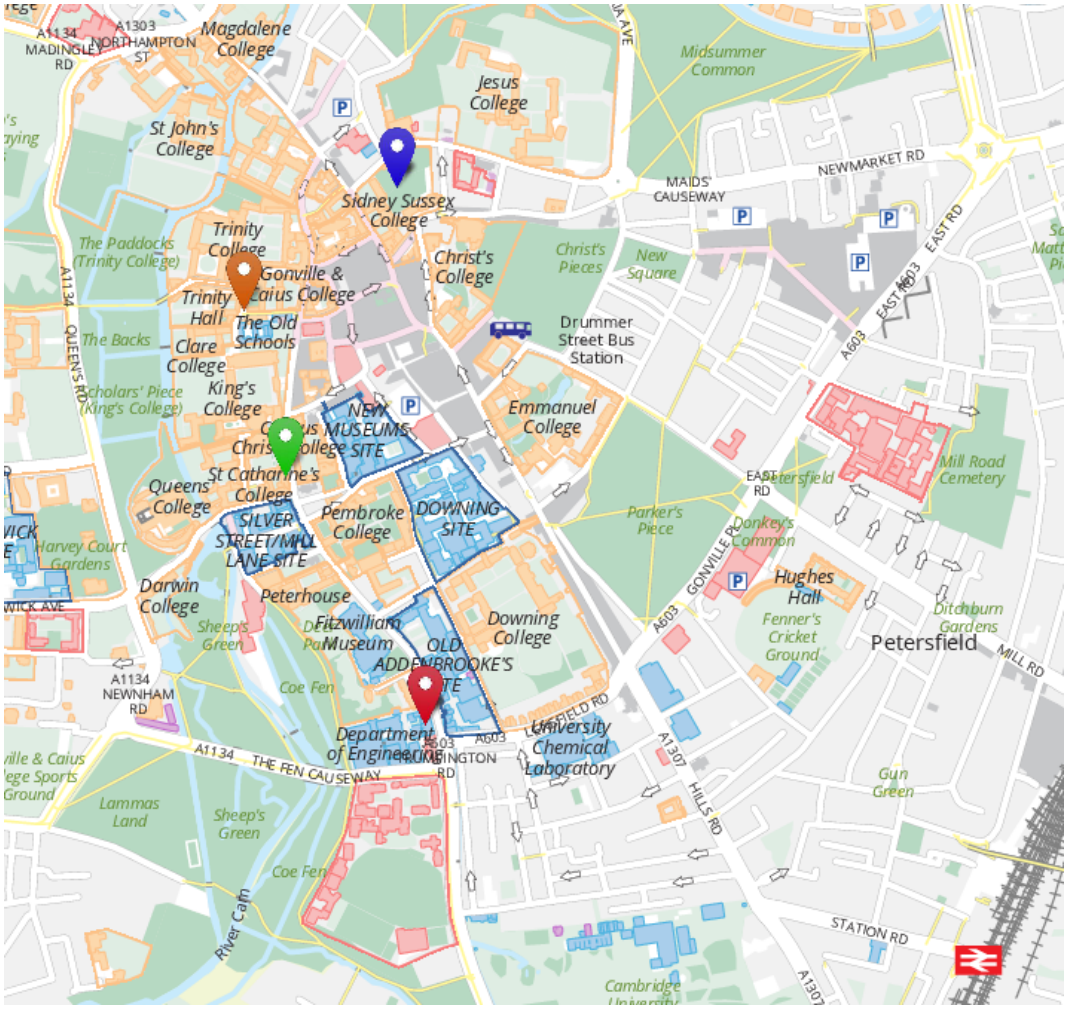
Monday September 11th

Time	Location	Session
11:30-12:15	CUED reception	Registration (continues until 13:30 for those not having lunch)
12.30-13:15	St. Catharine's College	Lunch (register at CUED in advance, 10 minutes walk)
13:30-13:45	CUED - LT1	Welcome
13:45-14:45	CUED - LT1	Keynote: *Acoustic Scene Mapping for Robot Audition* Christine Evers, Imperial College
14:50-15:20	CUED - LT1	Oral Session 1
15:20-15:35	CUED - LR3A	Break (tea and coffee)
15:35-16:50	CUED - LR3	Poster Session 1
16:50-17:00		Break
17:00-18:00	CUED - LT1	Oral Session 2
18:45-19:30	Trinity Hall	Drinks reception (sponsored by Apple)
19:30	Trinity Hall	Dinner (sponsored by Amazon)

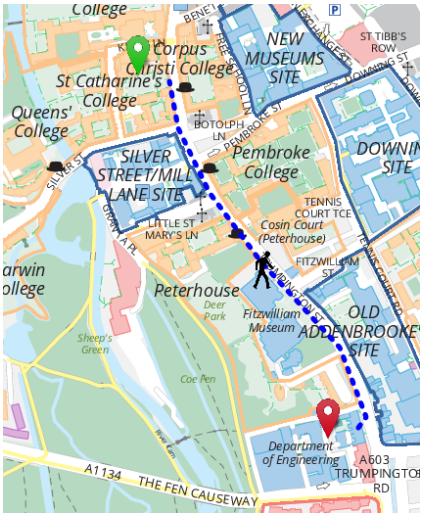
Tuesday September 12th

Time	Location	Session
08:40-09:00	CUED reception	Registration
09:00-09:30	CUED - LT1	Keynote: *Spoken Language Understanding in Alexa* Spyros Matsoukas, Amazon
09:30-10:45	CUED - LR3	Poster Session 2
10:45-11:00	CUED - LR3A	Break (tea and coffee)
11:00-12:15	CUED - LR3	Poster Session 3
12:15-13:45	St Catharine's College	Lunch
13:45-14:45	CUED - LT1	Keynote: *Tutorial on Sequence Models Using TensorFlow* Vincent Wan, Google
14:45-15:00	CUED - LT1	Final remarks and farewell

Map and Directions



- **Cambridge University Engineering Department (CUED)** - registration, talks and posters
- **St. Catharine's College** - lunch
- **Trinity Hall** - drinks reception and dinner
- **Sidney Sussex College** - accommodation



Social Events

On Monday 11th, there will a drinks reception and dinner at Trinity Hall. Trinity Hall dates back to 1350 and is one of the oldest and most beautiful of the Cambridge Colleges.

Bishop Bateman of Norwich originally founded Trinity Hall to promote the study of canon and civil law, probably due to the shortage of clergyman and lawyers following the Black Death of 1349. To this day, the College maintains a very strong tradition in the study of Law.

The dining hall, in which you will be enjoying dinner, is still housed in its original building dating from 1350. Updated in the Georgian period, it still has many features dating from its medieval beginnings, such as the stained glass windows and its minstrels gallery. The walls of the dining hall are lined with an impressive portrait collection including pictures of former Masters of Trinity Hall and benefactors of the College. The most recent addition to the dining hall is a portrait of the first two female Fellows at Trinity Hall Dr Kareen Thorne and Dr Sandra Raban.

Keynotes

Acoustic scene mapping for robot audition
Christine Evers, Imperial College, Monday September 11th, 13:45

Recent advances in robotics and autonomous systems are rapidly leading to the evolution of machines that assist humans across the industrial, healthcare, and social sectors. For intuitive interaction between humans and machines, spoken language is a fundamental prerequisite. However, in realistic environments, speech signals are typically distorted by reverberation, noise, and interference from competing sound sources. Acoustic signal processing is therefore necessary in order to provide machines with the ability to learn, adapt and react to stimuli in the acoustic environment. The processed, anechoic speech signals are naturally time varying due to fluctuations of air flow in the vocal tract. Furthermore, motion of a human talker s head and body lead to spatio-temporal variations in the source positions and orientation, and hence time-varying source-sensor geometries. Therefore, in order to listen in realistic, dynamic multi-talker environments, robots need to be equipped with signal processing algorithms that recognize and exploit constructively the spatial, spectral, and temporal variations in the recorded signals. Bayesian inference provides a principled framework for the incorporation of temporal models capturing prior knowledge of physical quantities, such as the acoustic channel. This talk therefore explores the theory and application of Bayesian learning for robot audition, addressing novel advances in acoustic Simultaneous Localization and Mapping (aSLAM), sound source localization and tracking.

Spoken language understanding in Alexa
Spyros Matsoukas, Amazon, Tuesday September 12th, 9:00

We will give an overview of Alexa's spoken language understanding components including wake-word detection, speech recognition, intent and named entity recognition, dialog management, and text-to-speech synthesis, and present a set of speech recognition and natural language understanding techniques we have developed as part of our continued efforts to enhance Alexa's conversational capabilities.

Tutorial on sequence models using TensorFlow
Vincent Wan, Google, Tuesday September 12th, 13:45

In this tutorial I'll be giving an in-depth introduction to the higher level TensorFlow APIs, including code snippets. I'll start from building simple graphs and running them. I'll cover feed forward neural networks, convolutional neural networks and finally show how to build a sequence to sequence model in TensorFlow. A basic knowledge of Python is assumed.

Technical programme

Oral Session 1
Monday 11th September, 14:50-15:20, LT1

1. M. Wester, M. P. Aylett, D. A. Braude: "Bot or not? Exploring the fine line between cyber and human identity"
2. A. Malinin, K. Knill, A. Ragni, Y. Wang, M. J. F. Gales: "An attention based model for off-topic spontaneous spoken response detection: An initial study"

Oral Session 2
Monday 11th September, 17:00-18:00, LT1

1. M. Isaac, E. Pfluegel, G. Hunter, J. Denholm-Price, D. Attanayake, G. Coter: "Improving automatic speech recognition for creation and editing of mathematical text through incremental parsing"
2. E. Loweimi, J. Barker, T. Hain: "Channel compensation in the generalised vector Taylor series approach to robust ASR"
3. C. Zhang and P. Woodland: "Tandem system joint training using discriminative sequence training for deep neural networks with a Gaussian mixture model output layer"

Bot or Not? Exploring the fine line between cyber and human identity

Mirjam Wester, Matthew P. Aylett, David A. Braude

CereProc Ltd., Edinburgh, UK

{mirjam,matthewa,dave}@cereproc.com

Abstract

Speech technology is rapidly entering the everyday through the large scale commercial impact of systems such as Apple Siri and Amazon Echo. Meanwhile technology that allows voice cloning, voice modification, speech recognition, speech analytics and expressive speech synthesis has changed dramatically over recent years. Using 'Bot or Not'[1] –an educational tool in the form of an online quiz that was designed to show-case and explain how technology of speech synthesis and voice modification has moved forward in recent years– we gathered impressions of what people realise is possible with current speech synthesis technology. The quiz consisted of 20 items, which all illustrated something relevant about speech synthesis. Information covered included: using found data to create voices, unit selection, speech styles, identity, emotion/prosody, phonetic coverage, voice modification, vocoding, statistical speech synthesis, cross-lingual speech synthesis. Feedback on 'Bot or Not' was collected in three different ways: by means of a focus group meeting, at the British Science Museum Lates, and via CrowdFlower, a crowd-sourcing platform.

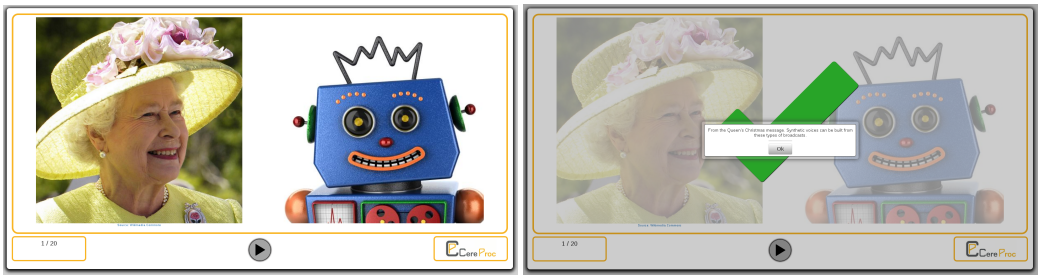


Figure 1: *Screen shots of the first item in the quiz. Left: an audio file is played, the user decides whether it is bot or not and selects the corresponding photo. Right: After clicking the photo, feedback is presented in a text box.*

In all three settings for collecting feedback, the reaction to the quiz was overwhelmingly positive. By first running a focus group meeting, the tool was effectively test-run prior to presentation at the Museum Lates and a short set of questions could be formulated to elicit more specific feedback. Although no explicit conclusions can be drawn from this experiment, it does highlight some of the ethical challenges that one must consider when including voice in an intelligent agent. Although human-like sounding synthesis was preferred over a robotic sounding voice, participants across the board want to know when they are hearing a synthetic voice, i.e., is it a bot or not?

[1] https://cerevoice.com/apps/botornot

An attention based model for off-topic spontaneous spoken response detection: An Initial Study

Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang and Mark J. F. Gales
Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{am969, mjfg}@eng.cam.ac.uk

August 18, 2017

Abstract

Automatic spoken language assessment systems are gaining popularity due to the rising demand for English second language learning. Current systems primarily assess fluency and pronunciation, rather than semantic content and relevance of a candidate's response to a prompt. However, to increase reliability and robustness, relevance assessment and off-topic response detection are desirable, particularly for spontaneous spoken responses to open-ended prompts. Previously proposed approaches usually require prompt-response pairs for all prompts. This limits flexibility as example responses are required whenever a new test prompt is introduced.

This work presents a initial study of an attention based neural model which assesses the relevance of prompt-response pairs without the need to see them in training. This model uses a bidirectional Recurrent Neural Network (BiRNN) embedding of the prompt to compute attention over the hidden states of a BiRNN embedding of the response. The resulting fixed-length embedding is fed into a binary classifier to predict relevance of the response. Due to a lack of off-topic responses, negative examples for both training and evaluation are created by randomly shuffling prompts and responses. On spontaneous spoken data this system is able to assess relevance to both seen and unseen prompts.

Improving Automatic Speech Recognition for Creation and Editing of Mathematical Text Through Incremental Parsing

Marina ISAAC[a], Eckhard PFLUEGEL[a], Gordon HUNTER[a],
James DENHOLM-PRICE[a], Dilaksha ATTANAYAKE[a] and Guillaume COTER[b]

[a] *Kingston University, London, U.K.*
[b] *Ecole des Mines de Douai, Douai, France*

Abstract

Activities involving creation and modification of mathematical content, such as equations and formulae in electronic documents, are becoming more and more important. Although facilities exist to do this using clicks or drags on the desktop, these do not provide a solution for people with disabilities involving motor control or use of their hands. One viable alternative input modality is automatic speech recognition (ASR), which has matured significantly over the last 20 years. The *TalkMaths* system aims to analyse spoken mathematical content supplied and transcribed by a standard ASR system in order to display this content correctly and permit editing thereof. Successive prototypes have achieved aspects of this.

The essential starting point for such analysis is the definition of a standard for the language used to describe the content in spoken form. Our application is influenced by the work of Fateman [1], who discusses the problem of ambiguity when reading mathematical expressions aloud. At the simplest level of arithmetic, ambiguity in mathematical content is resolved by rules of operator precedence, brackets and associativity. As more complexity is introduced, the lack of grouping constructs becomes apparent: for example, without these it would be impossible to dictate the expression $\sqrt{b^2 - a}$ without potential ambiguity: a mechanism is needed to allow users to specify such delimiters while minimising their cognitive load. The *TalkMaths* system addresses this issue by the use of templates, which range from simple bracketing structures (e.g. "square root"…"end square root" for the above example) to multi-component templates to permit specification of elements such as integrals.

The *TalkMaths* system processes mathematical expressions, transcribed from spoken form, to produce a parse tree that can be further manipulated. An attempt is made to identify and resolve errors in syntactically incorrect expressions. The first generation of *TalkMaths* used an attribute grammar, but this did not support incomplete input. Attanayake et al [2] introduced support for incomplete expressions, so as this task was not feasible using such a grammar, employed an operator precedence (OP) grammar to define the language. This was parsed using an OP parser adapted to handle mixfix (also known as distfix) operators to implement the templates, and incorporate error recovery strategies to handle incomplete input.

Although this approach is very efficient for the length of token stream that can be generated by a single utterance, we encounter a perceptible delay for longer expressions. This is because the entire input stream needs to be reparsed as the user dictates further after a pause. Incremental parsing, first developed decades ago to address similar problems for compiling computer code, provides a solution for this. As the language is expressed using an OP grammar, we may exploit the speed of incremental OP parsing algorithms. Our algorithm is an extension of that developed by Heeman [3] which effects editing actions using just two operations *merge* and *split* on the entire parse tree, combined with the node rearrangements developed by Lalonde & des Rivieres [4] and used by Kaiser & Kant [5] in their editor-centric approach. Most significantly, it handles Attanayake's composite template nodes by expanding on Heeman's simple bracket pair matching approach.

Performance evaluation is underway: simulations of batch parsing a new expression fragment and merging it with an existing parse tree far outperforms reparsing a concatenation of the input, in terms of re-parsing time; initial findings for a scenario involving the *split* operation also indicate performance using incremental parsing is significantly faster. In the future, we plan to evaluate use of this incremental parsing algorithm in a mobile implementation of *TalkMaths*, as well as to adapt it for use in an editor for computer program code.

References

[1] R. Fateman, "How can we speak math?," 2009. [Online]. Available: http://www.eecs.berkeley.edu/~fateman/papers/speakmath.pdf. [Accessed 03/08/2017].

[2] D. R. Attanayake, J. Denholm-Price, G. Hunter and E. Pfluegel, "TalkMaths over the web: a web-based speech interface to assist disabled people with mathematics," in *Proceedings of the Institute of Acoustics*, vol 36, part 3, pp. 435-442, 2014.

[3] F. C. Heeman, "Incremental parsing of expressions," *Journal of Systems and Software,* pp. 55-69, 1990.

[4] W. R. Lalonde and J. des Rivieres, "Handling Operator Precedence in Arithmetic Expressions with Tree," *ACM Trans. Program. Lang. Syst.*, vol. 3, pp. 83-103, 1981.

[5] G. E. Kaiser and E. Kant, "Incremental parsing without a parser," *Journal of Systems and Software,* pp. 121-144, 1985.

CHANNEL COMPENSATION IN THE GENERALISED VECTOR TAYLOR SERIES APPROACH TO ROBUST ASR

Erfan Loweimi, Jon Barker and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK

{{eloweimi1, j.p.barker,t.hain}@sheffield.ac.uk

ABSTRACT

Designing good normalisation to counter the effect of environmental distortions is one of the major challenges for automatic speech recognition (ASR). The Vector Taylor series (VTS) method is a powerful and mathematically well principled technique that can be applied to both feature and model domains to compensate for additive and channel noise. However, in its standard form, it is tied to MFCC (and log-filterbank) features and does not extend to other representations such as PLP, PNCC and phase-based front-ends that use power transformation rather than log compression. In [1], we aimed at broadening the scope of the VTS method by deriving a new formulation that assumes a power transformation is used as the nonlinearity during feature extraction. It is shown that the conventional VTS, in the log domain, is a special case of the new extended framework. In addition, the new formulation introduces one more degree of freedom which makes it possible to tune the algorithm to better fit the data to the statistical requirements of the ASR back-end. Performance-wise, generalised VTS (gVTS) provides performance improvement in both clean and additive noise conditions.

In [2] two more contributions were made. Firstly, while the previous gVTS formulation assumed that noise was purely additive, we now derive gVTS formulae for the case of speech in the presence of both additive noise and channel distortion. Second, a novel iterative method for estimating the channel distortion which utilises gVTS itself and converges after a few iterations is proposed. Since the new gVTS blindly assumes the existence of both additive noise and channel effects, it is important not to introduce extra distortion when either are absent. Experimental results conducted on CSR Aurora-4 database show that the new formulation passes this test. In the presence of channel noise only, it provides relative WER reductions of up to 30% and 26%, compared with previous gVTS and multi-style training with cepstral mean normalisation, respectively.

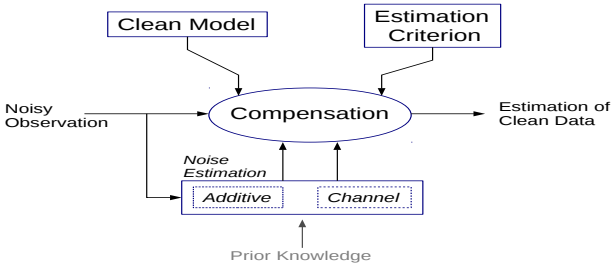


Fig. 1. Elements of the noise compensation process.

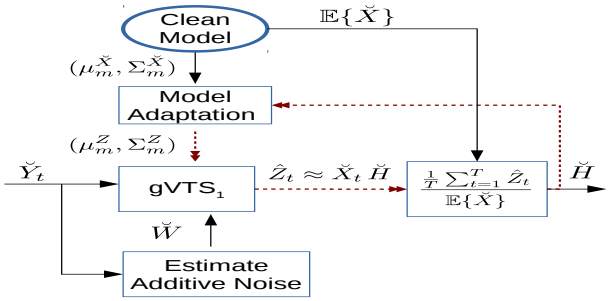


Fig. 2. Workflow of the proposed channel estimation method [2].

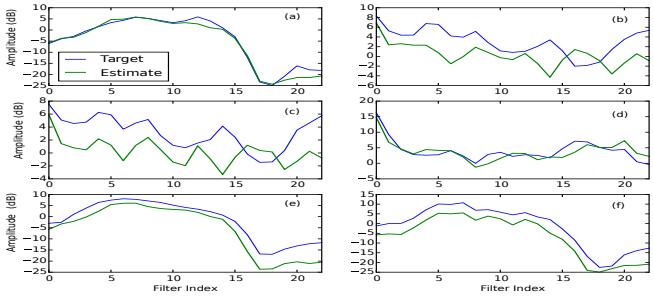


Fig. 3. Blind channel estimation based on the proposed method for 6 waves from the test set C of the Aurora-4. Target channel response was computed through comparing the noisy wave (Y) with its clean counterpart (X) from test set A. Underestimation is due to Jensen's inequality in (19) [2].

Table 1. *WER for Aurora-4 (HMMs trained on clean data) [2].*

Feature	A	B	C	D	Ave_1	Ave_2
MFCC-Clean	6.8	33.4	23.8	50.2	38.0	28.6
MFCC-Multi1	9.0	18.0	23.7	35.4	25.2	21.5
MFCC-Multi2	10.0	17.2	19.6	31.2	23.0	19.5
gVTS1-0.05	6.5	19.9	21.1	37.0	26.6	21.3
gVTS2-0.05-0	6.6	20.3	16.7	35.6	25.6	19.8
gVTS2-0.05-1	6.5	20.9	15.9	35.0	25.6	19.6
gVTS2-0.05-2	6.5	20.8	**14.4**	35.1	25.5	**19.2**
gVTS2-0.05-3	6.6	21.3	15.0	35.3	25.8	19.5
gVTS2-0.075-2	7.0	20.8	15.2	35.1	25.6	19.5
gVTS2-0.1-2	7.4	20.3	15.7	34.9	25.3	19.6

$Ave_1 = \frac{A+6B+C+6D}{14}$ $\quad$ $Ave_2 = \frac{A+B+C+D}{4}$

1. REFERENCES

[1] E. Loweimi, J. Barker, and T. Hain, "Use of generalised nonlinearity in vector taylor series noise compensation for robust speech recognition," in *Interspeech*, 2016, pp. 3798–3802.

[2] E. Loweimi, J. Barker, and T. Hain, "Channel compensation in the generalised vector taylor series approach to robust asr," in *Interspeech*, 2017.

Tandem system joint training using discriminative sequence training for deep neural networks with a Gaussian mixture model output layer

Chao Zhang and Phil Woodland

The use of deep neural networks (DNNs) for feature extraction and Gaussian mixture models (GMMs) for acoustic modelling is often termed a tandem system configuration and can be viewed as a DNN with a GMM output layer rather than an affine transform layer with a softmax activation function (DNN-GMM). Compared to the direct use of DNN output probabilities in the acoustic model, the tandem approach suffers from a major weakness in that the feature extraction stage and the final acoustic models are optimised separately. This paper proposes a joint optimisation approach to all the stages of the tandem acoustic model by using DNN-GMM discriminative sequence training. A set of techniques is used to improve the training performance and stability. Experiments using the multi-genre broadcast (MGB) English data show that the proposed method produced a 6% relative lower word error rate (WER) than that of a traditional discriminatively trained tandem system. The resulting jointly optimised tandem systems are comparable in WER to hybrid DNN systems optimised using discriminative sequence training with the same number of parameters.

Poster Session 1
Monday 11th September, 15:35-16:50, LR3

1. S. Renals: "The SUMMA media monitoring project"
2. F. L. Kreyssig, C. Zhang, P. C. Woodland: "Modular construction of complex deep learning architectures in HTK"
3. R. Errattahi, A. El Hannani, T. Hain, H. Ouahmane: "Efficacy of decoder versus non-decoder features for automatic speech recognition errors detection and classification: A comparative study"
4. A. Ragni, K. Knill, M. Gales, D. Kousoulidis: "Language independent bootstrapping for automatic speech recognition and keyword search in limited resource conditions"
5. C. Seivwright, M. Russell, S. Houghton: "Exploring a measure of discontinuity between adjacent trajectories in a continuous state HMM"
6. S. Deena, R. W. M. Ng, P. Madhyastha, L. Specia, T. Hain: "Semi-supervised adaptation of RNNLMs by fine-tuning with domain-specific auxiliary features"
7. K. M. Knill, M. J. F. Gales, K. Kyriakopoulos, A. Ragni, Y. Wang: "Use of graphemic lexicons for spoken language assessment"
8. M. Qian, X. Wei, P. Jančovič, M. Russell: "The University of Birmingham 2017 SLaTE CALL shared task systems"
9. L. Tian, J. D. Moore, C. Lai: "Recognizing emotions in spoken dialogue with acoustic and lexical cues"
10. W. A. Jassim, N. Harte: "On estimation of a priori SNR using neurograms for speech enhancement"
11. L. Baghai-Ravary, S. W. Beet: "Automatic telephone voice analysis and therapy"
12. J. Drayton, E. Miranda, A. Kirke: "Acoustic-articulatory parameter inversion of vowels by genetic algorithm and the Praat articulatory synthesizer"
13. S.-R. Lee, M. Huckvale: "Deep Neural Network (DNN) architecture for speech prediction"
14. B. R. Cowan, J. P. Cabral, L. Clark: "The CogSIS project: Examining the cognitive effects of speech interface synthesis"
15. M. S. Al-Radhi, T. G. Csapó, G. Németh: "Effects of adding a harmonic-to-noise ratio parameter to a continuous vocoder"
16. M. Wester, D. A. Braude, B. Potard, M. P. Aylett, F. Shaw: "Real-time reactive speech synthesis: incorporating interruptions"
17. H. L. Bear: "Visual gesture variability between continuous speech talkers"
18. T. Ash, N. S. Martínez-Santos, R. Braun, D. Mrva, M. Hollands, F. Skála, R. Francis, J. Gilmore, P. Hewlett, T. Robinson: "The Speechmatics Automatic Linguist (AL): Our platform for using ASR to build the languages of the world"

The SUMMA Media Monitoring Project

Steve Renals (University of Edinburgh) and the SUMMA Consortium

www.summa-project.eu

SUMMA (Scalable Understanding of Multilingual Media) () is a three-year EU H2020 *Research and Innovation Action* (Feb 2016 – Jan 2019) which is developing a highly scalable, integrated web-based platform to automatically monitor an arbitrarily large number of public broadcast and web-based news sources.

Two concrete use cases drive the project:

- **Monitoring of External News Coverage:** *BBC Monitoring* monitors about 15,000 distinct international sources, including some about 3,000 TV and radio channels. A single monitoring journalist typically monitors four TV channels and several online sources simultaneously. Automating the monitoring process will allow both broader (more sources) and deeper (richer analysis) coverage.
- **Monitoring Internal News Production:** *Deutsche Welle*, Germany's international public service broadcaster, provides international news and background information from a German perspective in 30 languages worldwide, and requires internal news monitoring in order to better share resources across production in multiple languages.

We are also exploring further use cases such as data journalism.

Achieving these goals requires advances in multilingual speech recognition, machine translation, metadata extraction from speech, and various natural language processing technologies inckuding knowledge base construction, automatic fact checking, summarisation, and sentiment analysis.

In this poster/demo, I'll give an overview of the SUMMA media monitoring platform, which is implemented as an orchestra of independent components that run as individual containers using Docker platform. This modular architecture allows the project partners to have a high level of independence in their development. The platform contains both a back-end data management and content processing system, and a front-end web interface. The demo will based on the initial version of the system which automatically processes live streams from BBC monitoring and Deutsche Welle (in English, Arabic, Russian, German, Spanish).

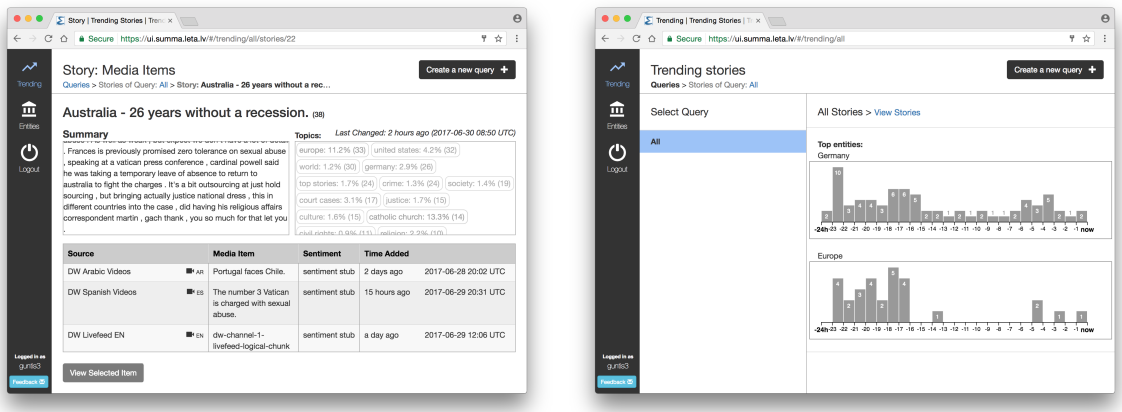


Figure 1: The SUMMA Platform

Modular Construction of Complex Deep Learning Architectures in HTK

F. L. Kreyssig , C. Zhang , and P. C. Woodland

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.
{flk24,cz277,pcw}@eng.cam.ac.uk

Abstract

In this work we present new extensions to the existing Artificial Neural Network (ANN) capabilities of the HTK speech recognition toolkit. These include new types of layers, different ways to combine input features as well as `PyHTK`, a python based library that provides a user-friendly interface through which one can set up complex deep learning architectures for ANN acoustic models.

`PyHTK` takes a simple config-file as its input that defines each layer of the network with its associated attributes. The attributes include the layer type, its input features and output dimension, its activation function and the type of operation that is applied to the input features before being fed into the layer. This operation can be a concatenation, as was the previous standard within HTK 3.5 [1], element-wise addition (ADD) or multiplication (MUL) of the inputs or the element-wise-applied MAX operation. For the training of recurrent architectures HTK currently uses unfolded training, which can take advantage of frame-level data shuffling. `PyHTK` automatically unfolds recurrent structures for a specified number of time-steps and also allows recurrent layers to operate at a lower frame rate as used in [2]. Bi-directional layers are also supported.

We evaluate a range of architectures on the TIMIT dataset, the results of which are shown in Table 1. The labels over 854 tied tri-phone states were derived from a set of 48 phone labels, which are mapped to the standard set of 39 phones for testing, after decoding on the full test set with a bigram phone language model. All models used 24 log-Mel filter bank coefficients with their Δ and $\Delta\Delta$ values as input features, except the CNN which used 40 without any Δ.

We create a ResNet-Block by appending two fully-connected (FC) layers with an "Activation-Only"-Layer that performs the ADD operation on the input to the first FC-layer and the output of the second, followed by the ReLU activation function. By combining nine blocks, preceded by one FC-layer, we create a 21-Layer ResNet, which achieves 20.37% PER. Using the SELU-activation function [3], we can train a 9-Layer MLP without pre-training or residual connections to a PER of 20.80%. Further, we also added LSTM, GRU and CNN layers to HTK. We test a CNN architecture with two convolutional layers followed by three fully connected layers, which achieves a PER of 20.12%. Deep-RNNs are easy to set up using the new interface. Stacking three ReLU-RNN layers yields a PER of 18.54%. Making the first two of these bi-directional gives an improvement to 18.15% PER. Stacking five RNN-layers achieved a performance of 17.48% PER. This network was pre-trained by training three and then four layers for one epoch each, with subsequent copying of the parameters. A noticeable improvement came from using ∼5x larger L2-regularization for the two pre-training epochs than for the succeeding epochs. All recurrent architectures used a 5-frame look-ahead buffer and were unfolded for 21 timesteps, thus having a total input context from `+5` to `-15`. All models were optimised using `SGD` with momentum.

Architecture	Width	PER
7L-RELU-MLP	500	21.43
9L-SELU-MLP	250	20.80
21L-(FC)ResNet	250	20.37
CNN	2048 for FC	20.12
3L-RELU-RNN	1024	18.54
3L-RELU-BDRNN	750	18.15
5L-RELU-RNN	750	17.48

Table 1: Phoneme error rates (PER) for the full TIMIT testset

References

[1] C. Zhang, P. C. Woodland, "A General Artificial Neural Network Extension for HTK.", in *Proc. Interspeech'15*.

[2] V. Peddinti and Y. Wang and D. Povey and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs", in *IEEE Signal Processing Letters*, 2017.

[3] G. Klambauer and T. Unterthiner, and A. Mayr and S. Hochreiter, "Self-Normalizing Neural Networks", *arXiv preprint arXiv:1706.02515*, 2017.

Efficacy of Decoder versus Non-Decoder Features for Automatic Speech Recognition Errors Detection and Classification: A Comparative Study

Rahhal Errattahi[1] *Asmaa El Hannani*[1] *Thomas Hain*[2] *Hassan Ouahmane*[1]

[1]*Laboratory of Information Technologies, National School of Applied Sciences*
University of Chouaib Doukkali, Morocco
[2]*Speech and Hearing Group, Department of Computer Science*
University of Sheffield, UK
{errattahi.r, elhannani.a, ouahmane.h}@ucd.ac.ma, t.hain@sheffield.ac.uk

1. Abstract

Automatic Speech Recognition (ASR) errors are essentially unavoidable. In order to detect errors in the output transcriptions of ASR systems, some previous work rely on critical information such as Word Lattices and other ASR decoder features. This information, which is not always available, highly depends on the decoding process. This poster addresses errors in continuous speech recognition output in two stages: the errors detection and the errors type classification, using different combination of features and different classifiers. Unlike the majority of research in this field, we propose to handle the speech errors independently to the ASR decoder using a set of contextual features derived exclusively from the ASR output.

Experiments were conducted using the Multi-Genre Broadcast (MGB) corpus [1] transcription output generated by the 2015 Sheffield's ASR system [2]. Two different types of feature set are investigated; decoder based features; the word confidence score, the confidence score of the previous word, and the confidence score of the next word, versus non-decoder based features; the LM probability of the word given its left context, the LM probability of the word given its right context, the sentence oddity [3] and the previous word label.

Results show that the SVM classifiers outperform both Neural Network and Bayesian Network classifiers. In addition, the SVM showed competitive performance in errors detection when using only non-decoder-based features as compared to decoder-based features. We also demonstrate that the same set-up could be applied to ASR error type classification task. We have confirmed that ASR errors are influenced by their context and that considering the label of the previous word in both tasks (i.e. errors detections and errors type classification) leads to better performance.

2. References

[1]. Bell P, Gales M, Hain T, Kilgour J, Lanchantin P, Liu X, McParland A, Renals S, Saz O, Webster M, et al (2015) The MGB challenge: Evaluating multi-genre broadcast media transcription. In: ASRU

[2]. Saz O, Doulaty M, Deena S, Milner R, Ng RW, Hasan M, Liu Y, Hain T (2015) The 2015 Sheffield system for transcription of Multi-Genre Broadcast Media. In: ASRU

[3]. Fong S, Skillicorn D, Roussinov D (2006) Detecting word substitution in adversarial communication. In: the proceedings of 6th SIAM Conference on Data Mining. Bethesda, Maryland

Language Independent Bootstrapping for Automatic Speech Recognition and Keyword Search in Limited Resource Conditions

Anton Ragni, Kate Knill, Mark Gales, Dimitris Kousoulidis
Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{ar527,kate.knill,mjfg,dk483}@eng.cam.ac.uk

August 17, 2017

Abstract

There are roughly 6,500 spoken languages in the world, the majority of which have had little or no attention within the speech processing community. To develop speech applications it is normally assumed that transcribed audio data is available for the language of interest. This paper examines to what extent it is possible to build, and train, automatic speech recognition (ASR) and keyword-spotting (KWS) systems when no target language transcribed data is available, but audio may be available. It is assumed that there are limited language model data, and a limited lexicon, available. To handle the lack of training data, a language-independent acoustic model is built, where a common phone set over all languages is used. This approach works reasonably well for some languages, but fails on others. To address this a bootstrap approach is adopted where audio data from the target language is recognised, and these hypotheses used to train an acoustic model. Furthermore, to address limitations of using limited language model data, the impact of text data from the web is investigated. Experiments with language-dependent and language-independent bootstrapped ASR and KWS systems are conducted on 4 IARPA Babel program languages including the surprise language of OpenKWS 2016 evaluation.

Exploring a Measure of Discontinuity Between Adjacent Trajectories in a Continuous State HMM.

Chloe Seivwright, Martin Russell, Steve Houghton

University of Birmingham, United Kingdom

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

1. Abstract

Motivated by the speech production process, this work implements a probabilistic trajectory segmental model [1] as a continuous state HMM (CS-HMM) [2], [3] . By using an underlying HMM model characterised by a continuous state space, you have the flexibility of enforcing continuity between two adjacent segments and therefore addressing the independence assumption that is implicit in a conventional HMM. Another positive feature of a CS-HMM is the compact decoding algorithm. By using a forward looking iterative method of decoding and exploiting the property that the product of two gaussian distributions will result in a scaled gaussian distribution, we have implemented an optimised parsimonious decoding system when compared to a computationally expensive segmental Viterbi decoder [4].

This work will present experimental results of two decoding systems, a piecewise linear decoder (PL-Decoder) where there is a discontinuity between adjacent segments, and a continuous piecewise linear decoder (PLC-Decoder) where you force continuity between adjacent segments, see Figure 1. A comparative HTK experiment shows compatible results when carefully limiting the number of parameters to match the CS-HMM.

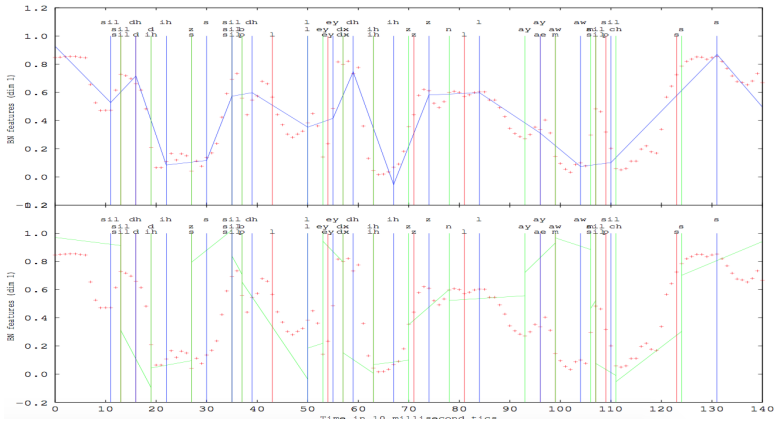


Figure 1: PLC Decoder (blue) and PL Decoder (green) output trajectories for a section of speech with TIMIT segmentation (red).

On analysing the output of these experiments via a statistical binomial significance test, it is possible to identify some interesting trends. The PL-Decoder outperforms the PLC-Decoder on the majority of consonant bursts and closures, whereas the PLC-Decoder performs better on *all* of the voiced phonemes. The results of this experiment is consistent with our prior understanding of speech, e.g. Abrupt changes in energy between consonants compared with smooth changes between vowels.

This result has encouraged us to explore the possibility of implementing a *soft* continuity constraint, i.e. a probabilistic measure of discontinuity between adjacent trajectories. These 'continuity measure' methods and results will be detailed and discussed.

2. References

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

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

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

[4] W. J. Holmes, "Modelling segmental variability for automatic speech recognition," Ph.D. dissertation, Department of Phonetics and Linguistics, University College London, 1997.

Semi-supervised Adaptation of RNNLMs by Fine-tuning with Domain-specific Auxiliary Features

Salil Deena[1], *Raymond W. M. Ng*[1], *Pranava Madhyastha*[2], *Lucia Specia*[2] *and Thomas Hain*[1]

[1]Speech and Hearing Research Group, The University of Sheffield, UK
[2]Natural Language Processing Research Group, The University of Sheffield, UK
{s.deena, wm.ng, p.madhyastha, l.specia, t.hain}@sheffield.ac.uk

1. Abstract

Recurrent neural network language models (RNNLMs) can be augmented with auxiliary features, which can provide an extra modality on top of the words. It has been found that RNNLMs perform best when trained on a large corpus of generic text and then fine-tuned sub-domain text for which it is to be applied. However, in many cases the auxiliary features are available for the sub-domain text but not for the generic text. In such cases, semi-supervised techniques can be used to infer such features for the generic text data such that the RNNLM can be trained and then fine-tuned on the available in-domain data with corresponding auxiliary features.

In this work, the features used for domain adaptation of RNNLMs are latent Dirichlet allocation (LDA) [1] features extracted from the text as in [2]. Whilst such features can be derived for the whole text, a setup is devised where the features are assumed to be available for the in-domain text but not for the generic text. In particular, the following semi-supervised adaptation approaches are proposed, aimed at compensating for the missing features:

1.1. Zero Features to Mask the Feature Sub-network

Zero features allows the training of a RNNLM with the same structure as the final RNNLM where features are available whilst masking the feature sub-network and therefore give an equivalent network to one without features. The advantage of this approach is it does not require structural changes to the network at the time of fine-tuning, but it does require the whole network to reconfigure.

1.2. Adding Feature Sub-network at Fine-tuning

This approach is similar in spirit to the zero features approach but can be convenient when an RNNLM has been trained already with possibly large amounts of data (and therefore requiring days to weeks to train) and then the network needs to be fine-tuned with in-domain text where features such as LDA are available. At the time of fine-tuning, the feature sub-network is introduced with the weights initialised with random values. The weights of the feature sub-network are allowed to update and thus reach a new convergence point with the rest of the network parameters.

1.3. Deriving the Features using Back-propagation

Here, a RNNLM is first trained with the show-level LDA features on the in-domain data. The network is then kept fixed and back-propagation is carried out on the generic text in order to recover the features f for each sentence. The features updated at each time step and the feature vector obtained at the end of the sentence is taken to be the sentence-level feature. The sentence-level features are then averaged at the show-level, to give the target show-level LDA features, which are then used to train a feature-based RNNLM on the generic text.

1.4. Deriving the Features using a Parametric Model

A parametric regression model is proposed, which can learn a mapping between the hidden state vector of a separate RNNLM trained without features, and the feature vectors. The hidden state vectors are first generated at sentence level by feeding each sentence through the RNNLM and extracting the hidden states at the end of the sentence. These hidden state vectors are averaged across all sentences in a given show, in order to give show-level hidden state vectors. A parametric regression model is then fitted between the averaged hidden state vectors and the show-level LDA features, which is then used to predict LDA features on the generic text.

2. Results

The main contribution of this work is showing that LDA features can be predicted for the generic text corpus, comprising about 80% of the whole data using the remaining 20%, giving a correlation of about 71% with ground truth LDA features. Both PPL and ASR results are also presented on Multi-Genre Broadcast data from the BBC [3], with results matching ground truth LDA features.

3. References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[2] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, "Combining Feature and Model-Based Adaptation of RNNLMs for Multi-Genre Broadcast Speech Recognition," in *Proc. of ISCA Interspeech*, 2016, pp. 2343–2347.

[3] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster, and P. Woodland, "The MGB challenge: Evaluating multi–genre broadcast media transcription," in *Proc. of IEEE ASRU*, 2015.

Use of Graphemic Lexicons for Spoken Language Assessment

K.M. Knill, M.J.F. Gales, K. Kyriakopoulos, A. Ragni, Y. Wang

ALTA Institute / Engineering Department
Cambridge University, UK
{kate.knill,mjfg,kk492,ar527,yw396}@eng.cam.ac.uk

Abstract

Automatic systems for practice and exams are essential to support the growing worldwide demand for learning English as an additional language. Assessment of spontaneous spoken English is, however, currently limited in scope due to the difficulty of achieving sufficient automatic speech recognition (ASR) accuracy. "Off-the-shelf" English ASR systems cannot model the exceptionally wide variety of accents, pronunications and recording conditions found in non-native learner data. Limited training data for different first languages (L1s), across all proficiency levels, often with (at most) crowd-sourced transcriptions, limits the performance of ASR systems trained on non-native English learner speech. This paper investigates whether the effect of one source of error in the system, lexical modelling, can be mitigated by using graphemic lexicons in place of phonetic lexicons based on native speaker pronunications. Graphemic-based English ASR is typically worse than phonetic-based due to the irregularity of English spelling-to-pronunciation but here lower word error rates are consistently observed with the graphemic ASR.

Test Set	Ph/Gr	Vit	CN
Gujarati	Ph	34.3	33.7
	Gr	33.3	32.5
	Ph $\oplus$ Gr	-	31.6
Mixed	Ph	48.6	47.5
	Gr	47.2	46.1
	Ph $\oplus$ Gr	-	44.2

Table 1: *Phonetic (Ph) and graphemic (Gr) trigram %WER with Viterbi (Vit) and confusion network (CN) lattice rescoring on evaluation sets consisting of Gujarati L1 and 6 mixed L1 English learners.*

The effect of using graphemes on automatic assessment is assessed on different grader feature sets: audio and fluency derived features, including some phonetic level features; and phone/grapheme distance features which capture a measure of pronunciation ability. As a candidate's proficiency improves, their pronunciation becomes more native, with commensurate reduction in strain to the listener caused by L1 effects [1]. Explicit features to represent pronunciation in the grader should therefore help assessment. To overcome issues with coping with the large variation in pronunciations and the lack of a native speaker reference, an approach based on distances between phones is presented here. Each phone is defined relative to the pronunciation of each of the other phones. The full set of phone-pair distances describes the speaker's overall accent.

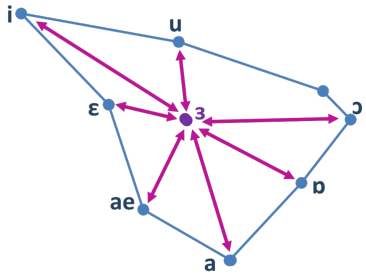


Figure 1: *Illustration of the phone distance concept*

Full details can be found in [2].

1. References

[1] C. of Europe, *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001.

[2] K. Knill, M. Gales, K. Kyriakopoulos, A. Ragni, and Y. Wang, "Use of Graphemic Lexicons for Spoken Language Assessment," in *Proceedings of INTERSPEECH*, 2017.

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.

The University of Birmingham 2017 SLaTE CALL Shared Task Systems

Mengjie Qian, Xizi Wei, Peter Jančovič, Martin Russell
School of Engineering, University of Birmingham, Birmingham B15 2TT, UK
{mxq486, xxw395, p.jancovic, m.j.russell}@bham.ac.uk

Abstract

The 2017 SLaTE CALL Shared Task [1] was led by the University of Geneva with support from the University of Birmingham and Radboud University using recordings of English responses from German speaking Swiss teenagers interacting with the CALL-SLT system [2]. The development set (ST-DEV) and the test set (ST-TST) consist of 5222 and 996 recordings respectively. We focused on the automatic speech recognition (ASR) part but also improved the text processing (TP) component. Our ASR system was developed using the Kaldi toolkit [3] and builds on the CALL Shared Task baseline ASR system. Several approaches to training a DNN-HMM ASR system using the AMI [4] and the German PF-STAR corpus [5], plus a limited amount of Shared Task data, are explored. In cross-validation evaluations on the initial Shared Task data, our final ASR system achieved a word error rate (WER) of 9.27%, compared with 14% for the official baseline Shared Task DNN-HMM system. For text processing we expanded the baseline template-based grammar to include additional correct response patterns from the original Shared Task transcriptions. Finally, we fused the outputs of several systems at the text processing stage using linear logistic regression.

For the final evaluation on ST-TST we submitted results from three systems:

- Submission 1 consists of our best ASR system trained on the whole of ST-DEV, plus the expanded TP. The optimal parameters of ASR for Submission 1 were estimated over 10-fold cross-validation experiments.
- Submission 2 is the result of fusing the outputs of six separate systems using linear logistic regression [6]. The systems all use our expanded TP with four variants of the ASR from Submission 1, the Kaldi baseline ASR and Nuance ASR.
- Submission 3 combines Nuance ASR with the expanded TP.

On ST-TST Submission 1 achieved a WER of 15.63% and Submission 1, 2 and 3 achieved D scores of 4.71, 4.766 and 2.533, respectively.

Reference

[1] C. Baur, J. Gerlach, E. Rayner, M. Russell, and H. Strik, “A Shared Task for Spoken CALL?” in Proc. Language Resources and Evaluation Conf. (LREC), 2016.

[2] C. Baur, “The Potential of Interactive Speech-Enabled CALL in the Swiss Education System: A large-Scale Experiment on the Basis of English CALL-SLT,” Ph.D. dissertation, Universitè de Genève, 2015.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, “The Kaldi speech recognition toolkit,” Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011.

[4] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal et al., “The AMI meeting corpus: A pre-announcement,” in Int. Workshop on Machine Learning for Multimodal Interaction. Springer, 2005, pp. 28–39.

[5] A. Batliner, M. Blomberg, S. D’Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, “The PF_STAR children’s speech corpus,” in Proc. European Conference on Speech Communication and Technology, 2005.

[6] N. Br¨ummer, “FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition Scores--Tutorial and user manual--,” Software available at http://sites. google.com/site/nikobrummer/focalmulticlass, 2007.

Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues

Leimin Tian, Johanna D. Moore, Catherine Lai
School of Informatics, the University of Edinburgh
s1219694@sms.ed.ac.uk, J.Moore@ed.ac.uk, clai@inf.ed.ac.uk

Introduction

Emotions play a vital role in human communications. Therefore, it is desirable for Human-Computer/Robot Interaction Systems to recognize and react to user's emotions during dialogue. However, state-of-the-art automatic emotion recognizers have limited performance compared to humans. Our work attempts to improve performance of recognizing emotions in spoken dialogue by identifying dialogue cues predictive of emotions, and by building multimodal recognition models with a knowledge-inspired hierarchical structure.

In particular, we propose features representing occurrences of DISfluency and Non-verbal Vocalisation (DIS-NV) in speech, and a HierarchicaL (HL) fusion strategy for building a multimodal recognition model that uses more abstract or global features at higher layers of its hierarchy. To study the effectiveness of the proposed approaches, we conduct emotion recognition experiments on two databases of English dialogue: the AVEC2012 database of spontaneous dialogue, and the IEMOCAP database of acted dialogue. The AVEC2012 database contains dyadic conversations between human subjects and Wizard-of-OZ style virtual agents, and annotated emotions of the human speakers. The IEMOCAP database contains conversations of pairs of actors either improvising instructed scenarios or acting pre-written scripts. Both databases annotate emotions on the dimension of Arousal (activeness), Expectancy (certainty), Power (dominance), and Valence (positive/negative). Our results show that including prior knowledge on emotions in dialogue in either the feature representation or the model structure is beneficial for automatic emotion recognition.

On Estimation of *a priori* SNR using Neurograms for Speech Enhancement

Wissam A. Jassim, Naomi Harte

Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

wissam.jassim@tcd.ie, nharte@tcd.ie

1. Abstract

In statistical-based speech enhancement algorithms, the challenge is to find a non-linear estimator using the set of DFT coefficients of noisy speech to estimate the corresponding unknown clean speech coefficients. Most of these algorithms require the *a priori* SNR $\xi(k)$ to be estimated. It is defined as the true SNR (before adding noise) of the kth spectral component of the signal. Most of the existing algorithms to estimate $\xi(k)$ were derived using acoustic-signal-based features. The most common algorithm is the decision-directed approach [1], based on the relationship between $\xi(k)$ and *a posteriori* SNR, $\gamma(k)$. The a posteriori SNR represents the measured SNR of the kth spectral component of the noisy signal. Several methods have been proposed to improve the performance of the decision-directed approach. For example, a statistical model that takes into account the time-correlation between successive spectral components of the speech signal was proposed in [2]. Also, a large bias in the decision-directed approach was located in [3]. This biasing may lead to an unwanted distortion in speech when the weight factor approaches one.

Motivated by the fact that neural responses exhibit strong robustness to noise, this study proposes a different approach to estimate the a priori SNR using neural-response-based features. A computational model of the auditory-nerve (AN) system by Zilany *et al.* [4] [5] was employed to simulate the neural responses. The input to the model is a speech stimuli and the output is the time-varying spike counts for AN fibers tuned to different characteristic frequencies (CFs) as a function of time. The outputs of the AN model corresponding to a range of CF values construct the *neurogram* which is a 2D representation (time-frequency). Features are extracted from the resultant 2D neurogram using the Radon transform to compute the projections of neurogram matrix along specified directions (rotation angles). The Radon features are combined with the corresponding speech features of each frame (magnitude of noisy speech, noise power spectral density, and the a posteriori SNR). The extracted feature sets were trained by Support Vector Regression (SVR) algorithm to predict the a priori SNR components of unseen frames. The performance of the proposed method was tested using the NOIZEUS database. A range of speech quality and intelligibility measures were used for performance evaluation. The results demonstrate that the estimate of the a priori SNR using neural features can improve the output of a range of speech enhancement algorithms, including MMSE, Wiener filtering and perceptually-motivated approaches.

2. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[2] I. Cohen, "Relaxed statistical model for speech enhancement and a priori snr estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[3] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Communication*, vol. 49, no. 78, pp. 530 – 541, 2007, speech Enhancement.

[4] M. S. Zilany and I. C. Bruce, "Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 402–417, 2007.

[5] M. S. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, 2009.

Automatic Telephone Voice Analysis and Therapy

Ladan Baghai-Ravary and Steve W. Beet

Advanced Speech Processing Team,
Aculab plc, Bramley Rd., Milton Keynes, MK1 1PT, United Kingdom

Abstract

We have developed a wide range of speech technology for use in telephony applications ranging from Automatic Speech Recognition (ASR) and Speaker Verification (SV) to voice manipulation, telehealth monitoring and, most recently, voice therapy. Together, our algorithms [1] capture many diverse aspects of a speaker and their speech, and do so in a way which is robust to most of the distortions prevalent in modern-day telephone communication. In particular, our estimates of pitch, jitter and noise-to-harmonic ratio (NHR) are almost completely unaffected by telephone transmission.

This paper will present an example of how such telephone-optimised measurements can be combined into a system which is simultaneously easy and intuitive for the end-user, effective, economical, scalable, and simple to maintain.

The pilot system that we will describe allows a Speech and Language Therapist (SLT) to set exercises to be performed by their clients, at a time and at a pace to suit the client. It uses a combination of DTMF (telephone key-press) detection, pitch analysis, and voice manipulation to provide a rapid and flexible dialogue, allowing the client to practice their voice control over any telephone, with no need to purchase any particular hardware or software.

Additional parameters can be used to detect anomalies within a call, whether due to non-compliance by the client, network problems resulting in packet loss, or other more severe distortions.

References

[1] “VoiScan: Telephone Voice Analysis for Health and Biometric Applications”, L. Baghai-Ravary and S. W. Beet, to appear in “Speech and Computer”, Springer, 2017

Acoustic–Articulatory Parameter Inversion of Vowels by Genetic Algorithm and the Praat Articulatory Synthesizer

Jared Drayton, Eduardo Miranda, Alexis Kirke

Interdisciplinary Centre for Computer Music Research

Abstract

Acoustic–Articulatory Inversion (AAI) has been a long standing problem within the field speech processing [1]. The ability to derive articulatory information production from only the speech signal would be particularly beneficial to a number of sub areas, for instance speech synthesis, speech coding, speech recognition, etc. The techniques for AAI can be roughly divided into four categories, as shown in Figure 1.

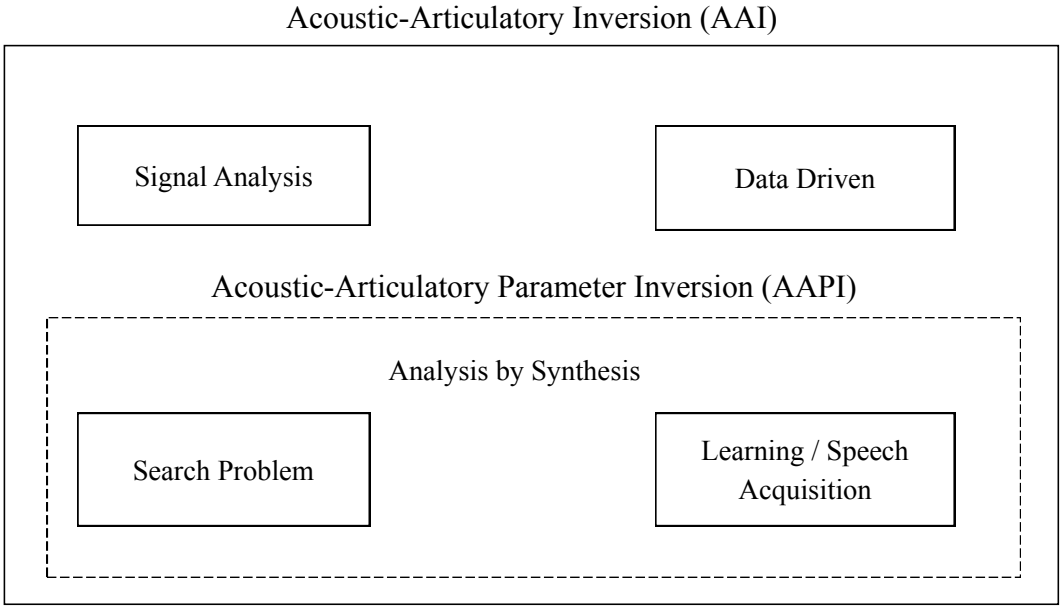


Figure 1: A taxonomy of Acoustic-Articulatory Inversion.

The last two can be be considered a subset of AAI, described here as Acoustic-Articulatory Parameter Inversion (AAPI), where an articulatory synthesizer is harnessed in an Analysis–by–Synthesis set up, where the aim is to find parameters for the synthesizer that allow it to reproduce a given target sound. The research in this poster approaches the problem as a search/optimization task, making use of Evolutionary Algorithm techniques and the Praat articulatory synthesizer. Results obtained from the latest experimentation with formant based fitness functions designed to be used with a Genetic Algorithm are presented.

References

[1] J. Schroeter and M. M. Sondhi, “Techniques for estimating vocal-tract shapes from the speech signal,” IEEE Trans. Speech Audio Process., vol. 2, no. 1, pp. 133–150, 1994.

Title: Deep Neural Network (DNN) Architecture for Speech Prediction
Authors: Sae-Ryoung Lee and Mark Huckvale
Affiliation: University College London, UK
Abstract

The advancement in speech technology has allowed us to imagine a talking machine with acceptable recognition accuracy, reaction time, and naturalness in style. The current research project stems from the question of what it takes for such a machine to learn to talk. Specifically, we explore whether the learning machine could recognise the statistical properties of phonetic and phonological patterns in order to generate speech-like output. The project motivates an answer to the question by building a simple neural-network-based acoustic model for speech synthesis, called a single-layer autoencoder. It consists of eight network parameters, only three of which is selected to investigate possible causal links. The idea is implemented using the CMU arctic slt speech database, on which the autoencoder is trained and tested. The predictive performance of the network is evaluated by two measures, including an objective reconstruction error and entropy/bit-rate (Kominek & Black, 2003). Then, a more complicated, stacked autoencoder (SAE) is constructed to investigate the effect of downsampling, which exploits the hierarchical property of the architecture in speech synthesis. The downsampling option allows us to compare the error rates between a single-layer autoencoder and a multi-layer autoencoder given the same baseline configuration. The network performance is further evaluated by a novel autopredictor program, which resembles an autoregressive model that 'fantasises' future output frames based on past input frames (Huckvale, 2017).

In our particular application, we found that the 'best' baseline configuration consisted of the bottleneck size of 40 units, the sparsity level of 20%, and the input frame size at 5-ms interval. Results from one-way ANOVA comparing the four acoustic models showed that depth of network was a reliable predictor for both reconstruction error and entropy. Fantasy output was classified into five different categories in the order of realistic speech sounds, including steady-state, silence, repetitive, constant non-speech, and babble-like sounds.

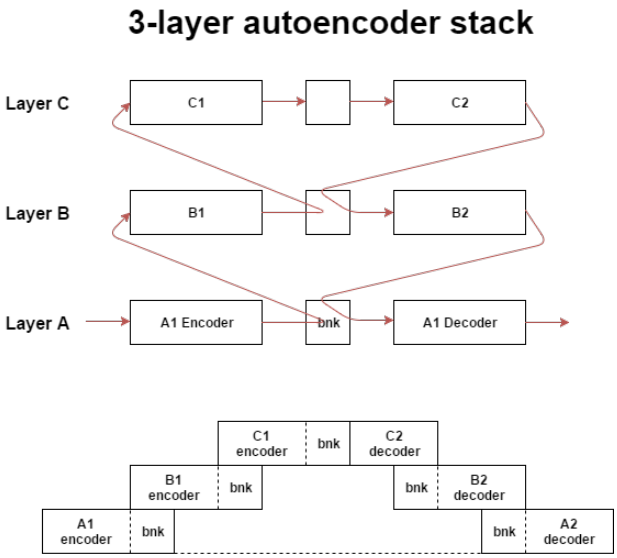


Fig 3. A three-layer stacked autoencoder. Each layer consists of an encoder, a bottleneck layer, and a decoder. In a stacked autoencoder, input for the successive layer comes from the previous bottleneck layer. The training order is indicated by red arrows.

The CogSIS Project: Examining the Cognitive Effects of Speech Interface Synthesis

Benjamin R. Cowan[1], João P. Cabral[2], Leigh Clark[1]

[1]University College Dublin, Ireland

[2]Trinity College Dublin, Ireland

1. Abstract

With the growth of intelligent personal assistants, pervasive and wearable computing and robot based technologies, speech interfaces are set to become a common dialogue partner. Technological challenges around the production of natural synthetic voices have been widely researched, yet comparatively little is understood about how synthesis affects user experience. The CogSIS project examines the psychological consequences of synthesis design decisions on the relationship between humans and speech technology. In particular we look to explore how design decisions around naturalness impact the assumptions we make about speech interfaces as communicative actors (i.e. our partner models). It fuses knowledge, concepts and methods from psycholinguistics, experimental psychology, human-computer interaction and speech technology to 1) understand how synthesis design choices, specifically accent and expressivity, impact a user's partner model, 2) how these choices interact with context to impact partner models and user experience 3) how these choices impact language production. The project will lead to a set of practical and actionable guidelines for synthesis design whilst also influencing synthesis evaluation practice.

Effects of adding a Harmonic-to-Noise Ratio parameter to a Continuous vocoder

Mohammed Salah Al-Radhi[1], *Tamás Gábor Csapó*[1,2], *Géza Németh*[1]

[1]Department of Telecommunication and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
{malradhi,csapot,nemeth}@tmit.bme.hu

1. Abstract

In this paper, we present an extension of a novel continuous residual-based vocoder for statistical parametric speech synthesis [1]. Our previous work has shown the advantages of adding envelope modulated noise to the voiced excitation [2]. The objective of this research is to show the effect of adding Harmonic-to-Noise Ratio (HNR) as a new excitation parameter to the Continuous vocoder. Experimental results based on objective evaluation have proven that the voice built with the proposed framework gives state-of-the-art speech synthesis performance while outperforming the previous baseline.

2. Experiment and evaluation

The general workflow of the proposed method is showed in Figure 1. For each frame, a fundamental frequency (F0), maximum voiced frequency (MVF), spectral envelope, and HNR [3] were estimated and recorded as main parameters for the Continuous vocoder. For purely harmonic signals, the HNR is positive infinite; while for the noise, it is negative infinite. Two English speakers were chosen from CMU-ARCTIC database, denoted AWB (Scottish English, male) and SLT (American English, female). 100 sentences (sampled at 16 kHz) from each speaker were analyzed and synthesized with the baseline and proposed vocoders. Based on mean Phase Distortion Deviation (PDD), the proposed method is closer to natural speech than the baseline (see Figure 2), which thereby encourages further studies.

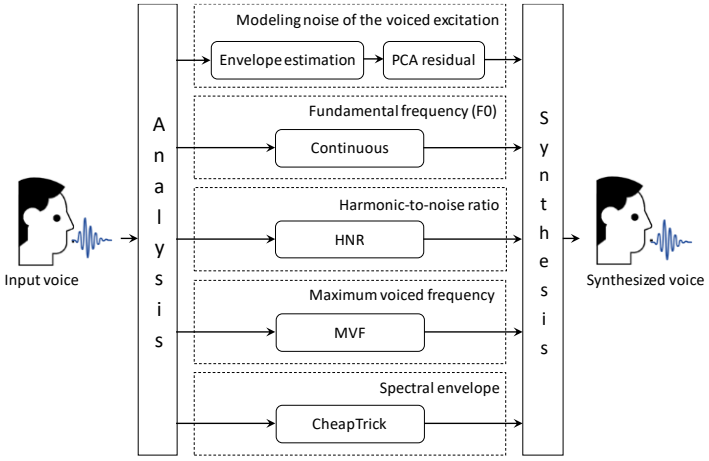


Figure 1: Workflow of the proposed method.

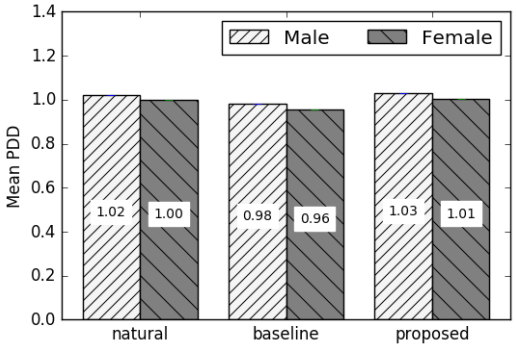


Figure 2: Mean PDD values by sentence type.

3. References

[1] Tamás Gábor Csapó, Géza Németh, Milos Cernak, and Philip N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in *EUSIPCO*, Budapest, 2016.

[2] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," in *Interspeech*, Stockholm, 2017.

[3] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *proceedings of the Institute of Phonetic Sciences*, Netherlands: University of Amsterdam, 1993.

Real-time reactive speech synthesis: incorporating interruptions

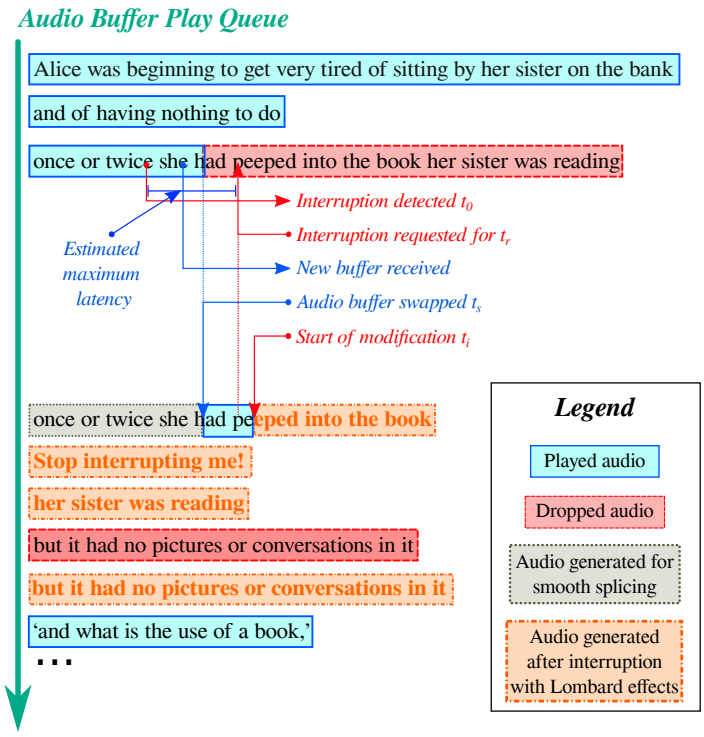
Mirjam Wester, David A. Braude, Blaise Potard, Matthew P. Aylett, Francesca Shaw

CereProc Ltd., Edinburgh, UK

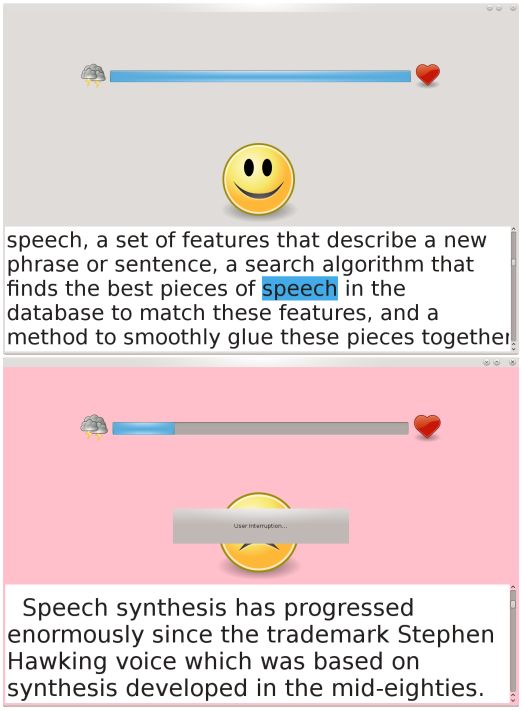
{mirjam,dave,blaise,matthewa,francesca}@cereproc.com

Abstract

The ability to be interrupted and react in a realistic manner is a key requirement for interactive speech interfaces. While previous systems have long implemented techniques such as 'barge in' where speech output can be halted at word or phrase boundaries, less work has explored how to mimic human speech output responses to real-time events like interruptions which require a reaction from the system. Unlike previous work which has focused on incremental production, here we explore a novel re-planning approach. The proposed system is versatile and offers a large range of possible ways to react. A focus group was used to evaluate the approach, where participants interacted with a system reading out a text. The system would react to audio interruptions, either with no reactions, passive reactions, or active negative reactions (i.e. getting increasingly irritated). Participants preferred a reactive system.



(a) *Example of the use of the interruption API, showing the changes in audio buffers. Final played audio is in blue and orange boxes, red and grey boxes are dropped. Note that $t_r - t_0$ must be larger than the maximum system latency.*



(b) *Visualisation of the system in action. The bar at the top and the emoji indicate the system's mood. The text being read is displayed at the bottom, with the current word being read highlighted. When an interruption occurs, the background colour is modified and a dialogue window is displayed.*

We discuss the relationship between incremental processing, re-planning and splicing. Incremental speech synthesis can be seen as a way of speeding up processing, but is not sufficient to achieve a reactive system. This leads to the following design guidelines for a reactive speech synthesis system:

1. **Fast enough** Whatever the chunk is (utterance, phrase, etc) the system must be able to synthesise replacement chunk within the required latency (we would suggest 200 ms as a minimum). For example, in our system where the chunk is a spurt (or intonational phrase) and 95% are less than 1850 ms for a 200 ms latency the system has to be at least 10x real-time.
2. **Splice audio in** You need to be able to tightly control audio output, i.e. be able to alter queued audio while it is waiting to be played and to know almost exactly what audio has been played. For multi-modal systems this would extend to the video output as well.
3. **Know how to respond** The appropriate response to an interruption varies considerably by application, as we found in our focus group, helpful systems should pause politely and rephrase, however for virtual characters a whole set of human responses to interruptions including rudely continuing may need to be implemented.

Visual gesture variability between continuous speech talkers

Helen L Bear {helen@uel.ac.uk}
University of East London, London, E16 2RD, UK.

In machine lip-reading, one attempts to interpret words from lip motions. Thus there are two 'translations' within the process, visemes into phonemes into words. However, there is no known map between phonemes to visemes. There is an argument for speaker-dependent maps between visemes and phonemes, but we need to know how dependent should these maps be? We try to determine, how different are speaker-dependent visemes?

We use a phoneme clustering algorithm to produce a series of speaker-dependent P2V maps of:

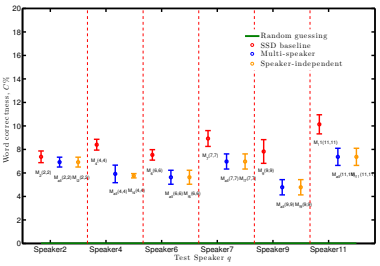
1. a multi-speaker P2V map using *all* speakers' phoneme confusions;
2. a speaker-independent P2V map for each speaker using confusions of all *other* speakers in the data;
3. a speaker-dependent P2V map for each speaker.

The 25 P2V maps are tested with separate training and test data over ten fold cross-validation. We use the HTK toolkit and report word correctness, C_w, to normalise for training sample bias. We track 12 RMAV speakers and extract AAM features. Each test in this experiment is designated as: $M_n(p,q)$ where P2V map M_n from speaker n, is trained using visual speech data from speaker p and tested using speaker q.

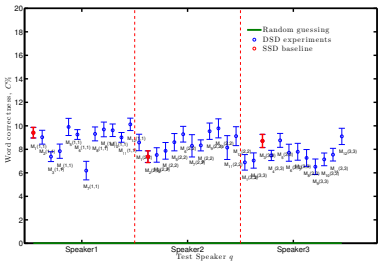
Our **MS** map is: $M_{[all]}(p,q)$ where $p=q$. Our **SI** tests use 12 maps derived using all speakers confusions bar the test speaker. '!n' means 'not including speaker n'. The SI maps are $M_{!1}(1,1)$, $M_{!2}(2,2)$ and so on. **Different Speaker-Dependent maps & Data** (DSD&D) tests use the maps and training data of the non-test speaker (see left of Table) and **Different Speaker-Dependent** maps (DSD) tests are the same speaker-dependent maps but now we train and test on the same speaker (see right of Table). Speaker dependent lip-reading benchmark (SSD) are speaker-dependent P2V results where $n=p=q$. This is our benchmark.

DSD&D				DSD			
Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$	Mapping (M_n)	Training data (p)	Test speaker (q)	$M_n(p,q)$
Sp2	Sp2	Sp1	$M_2(2,1)$	Sp2	Sp1	Sp1	$M_2(1,1)$
Sp...	Sp...	Sp1	$M_{..}(...,1)$	Sp...	Sp1	Sp1	$M_{..}(1,1)$
Sp12	Sp12	Sp1	$M_{12}(12,1)$	Sp12	Sp1	Sp1	$M_{12}(1,1)$

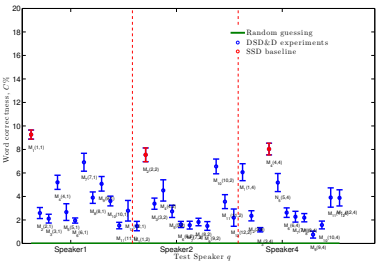
All **results** are measured in C_w% with error bars. In Figure (center) we see the effects of the unit selection. Figure (right) it is reassuring to see some speakers significantly deteriorate the classification rates when the speaker used to train the classifier is not the same as the test speaker as this is consistent with our MS and SI tests. If we compare these figures to the isolated words results, either the extra data in this larger data set or the longer sentences in continuous speech have made a difference. Whilst the lack of training on a test speaker is still significantly less accurate than speaker dependent tests, with continuous speech, it is significantly better than the difference experienced in isolated words.



MS and SI tests



DSD tests



DSD&D tests

We **conclude** there is high risk of over-generalising a MS/SI P2V map. These results suggest that a set of certain MS visemes with some SD visemes, may improve speaker-independent lip-reading. We have shown exceptions where the P2V map selection is significant and where classifiers trained on non-test speakers has not been detrimental. This gives hope that with visually similar speakers, speaker independent lip-reading is possible. Furthermore, with continuous speech speaker dependent P2V maps improve lipreading over isolated words. We attribute this to the co-articulation effects on phoneme recognition confusions which in turn influences the speaker-dependent maps with linguistic or context information. This is supported by evidence from conventional lipreading systems which show the strength of language models in lipreading accuracy.

The Speechmatics Automatic Linguist (AL): Our platform for using ASR to build the languages of the world

Tom Ash, Nicolás Serrano Martínez-Santos, Rudolf Braun, David Mrva, Michel Hollands, František Skála, Rémi Francis, James Gilmore, Paul Hewlett, Tony Robinson
Speechmatics, Cambridge, United Kingdom
{toma,tonyr}@Speechmatics.com

Abstract

The Automatic Linguist is our name for Speechmatics' platform to build automatic speech recognition language packs and adapt those language packs to specific domains.
It has the potential to build *any* language for which we have data to learn from.
Some languages have idiosyncratic features, in which case the performance may not be immediately optimal; as we encounter them we work to improve our performance on languages with these features.
All machine learning needs data to learn from; the Automatic Linguist can be trained on any of: commercial speech corpora, internally built corpora and / or customer supplied data.
The time to build depends very much on what data is available, but it may be as little as two weeks.

Using the Automatic Linguist we have automatically built ASR systems in nearly 30 of the world's most widely spoken languages, without employing linguistic natives and the resultant WERs are significantly better than more highly resources alternative commercial ASR providers.

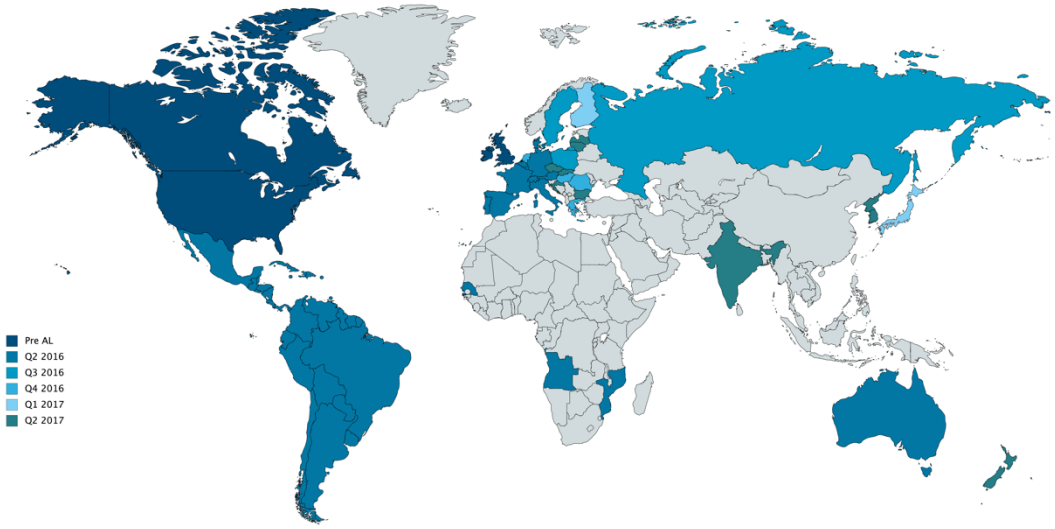


Figure 1: Map of the world, showing countries for which we have created an ASR system and when we created it. Most African countries have been omitted at present as although we have produced language packs in, for example, French, we have not tested on African accented data so are not certain of performance in those regions of the world.

Language	Speechmatics WER	Alternate Provider WER
English (US)	12.4%	21.3%
Spanish (Spain)	15.1%	27.1%
Hindi	23.3%	29.2%
Japanese	20.9%	27.8%

Figure 2: Word Error Rates on select languages when compared to a prestigious alternate provider of commercial ASR. Test sets were 4 hours long of data taken from realistic use cases - broadcast news, meeting recordings and call centre recordings.

Poster Session 2
Tuesday 12th September, 09:30-10:45, LR3

1. A. Haider, P. C. Woodland: "Comparison between Hessian free and natural gradient training for DNN acoustic models"
2. M. K. Baskar, M. Karafiát, L. Burget: "Residual memory networks: Feed-forward approach to learn long temporal dependencies"
3. J. H. M. Wong, M. J. F. Gales: "ASR teacher-student training and ensemble target diversity"
4. J. Rownicka, S. Renals, P. Bell: "Simplifying very deep convolutional neural network architectures for robust speech recognition"
5. X. Chen: "Future word contexts in neural network language models"
6. P. Bell, O. Klejch, S. Renals: "Automatic speech recognition for media monitoring in SUMMA"
7. Y. Wang, A. Ragni, X. Chen, K. M. Knill, M. J. F. Gales: "Improving automatic speech recognition for spoken language assessment"
8. Z. Hodari, S. King: "A learned emotion space for emotion recognition and emotive speech synthesis"
9. R. Nayar, M. Dunlop, A. Lowit, J. Soraghan, J. Levine: "Fine tuning neural network based phoneme recognition system for children's speech therapy"
10. X. Wei, M. Russell, P. Jancovic: "Automatic analysis of motivational interviewing with diabetes patients"
11. S. Ultes, L. Rojas-Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrksic, T.-H. Wen, M. Gasic, S. Young: "PyDial: A multi-domain statistical dialogue system toolkit"
12. J. Liu, F. Alegre, B. Fauve, A. Kanervisto, M. Sahidullah, T. Kinnunen, H. Delgado, M. Todisco, N. Evans, M. Falcone: "From media fears to research reality: How ready are countermeasures against speaker verification spoofing?"
13. J. P. Cabral, B. R. Cowan, K. Zibrek, R. McDonnell: "Evaluation of speech synthesis for virtual characters"
14. S. Ronanki, O. Watts, S. King: "A hierarchical encoder-decoder model for SPSS"
15. M. Wan, G. Degottex, M. J. F. Gales: "Integrated speaker-adaptive speech synthesis"
16. H. L. Bear, S. L. Taylor: "Visual speech recognition: aligning terminologies"

Comparison between Hessian Free and Natural Gradient training for DNN acoustic models

Adnan Haider, Philip C. Woodland

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K

mah90@cam.ac.uk, pcw@eng.cam.ac.uk

1. Abstract

In Automated Speech Recognition (ASR), DNN acoustic models are often trained using discriminative sequence training, that better approximates the Word Error Rate (WER) than frame-based training. Sequence training of DNNs is normally implemented using Stochastic Gradient Descent (SGD) or Hessian Free (HF) training. Second order methods like HF have shown to alleviate the problems of vanishing and exploding gradients by rescaling the gradient direction to adjust for the high non-linearity and ill-conditioning of the objective function. In a distributed setting where the gradient calculations can be easily parallelised, such an approach has shown to be time efficient [6].

An alternative approach to training DNNs that has recently experienced renewed popularity is the method of Natural Gradient (NG) [3, 4, 5]. This method was first proposed by Amari [1], as an effective optimisation method for training parametric density models. Instead of formulating gradient descent in the Euclidean parameter space, we establish a geometry on the space of density functions and perform steepest descent on that space. This work compares HF with an alternative batch style optimisation framework which employs NG to traverse through the parameter space. By correcting the gradient according to the local curvature of the KL-divergence using Conjugate Gradients (CG), the NG optimisation process can be shown to achieve greater improvements in both training speed and convergence by following a path in the parameter space where minimising the MPE or sMBR criterion correlates well with achieving reductions in WER. In ASR, NG has previously been applied for frame-based classification using small batches under a different optimisation framework. In [2], the author assumes a block diagonal structure for the empirical Fisher Information(FI) matrix and compute the NG direction explicitly. By contrast, this work makes no such assumptions about the structure of the empirical FI matrix and computes an approximation to the NG direction by running few iterations of CG. Unlike the framework described in [2], the optimisation framework discussed here has the flexibility that it can be used to train DNNs with respect to any sequence classification criterion.

The comparison between different optimisers was made by training models using a 200hr training set from the Multi-Genre Broadcast 1 (MGB) transcription task. As an evaluation set, we took the remaining 35 shows from the MGB1 dev.full that did not contain utterances from the official dev.sub set. We denote this set as dev.sub2. In this work, the model architecture of the DNNs comprised of 5 hidden layers each with 1000 nodes with sigmoid activation functions. The output softmax layer had context dependent phone targets formed by conventional decision tree context dependent state tying and contained 6000 nodes. Table 1 shows the WER results achieved on dev.sub2 with models using the 158k trigram model.

	CE	MPE			
		SGD	HF	DSAG-HF	NG
WER	33.1	30.8	31.0	30.8	30.5

Table 1: %WER differences on the MGB1 dev.sub2 set between different optimisers with 158k trigram LM

From our experiments with the MGB1 datasets, we have found that the batch-styled NG achieves the greatest WER reductions while possessing the same advantages as large batch Hessian free methods in terms of stability and being inherently data parallel.

2. References

[1] S. Amari, "Natural Gradient Works Efficiently in Learning," *Proc. Neural Information Processing Systems (NIPS)*, 1998.

[2] D. Povey, X. Zhang & S. Khudanpur, *Parallel Training of DNNs with Natural Gradient and Parameter Averaging*, arXiv preprint arXiv:1410.7455, oct 2014.

[3] R. Pascanu &Y. Bengio, *Revisiting Natural Gradient for Deep Networks*, arXiv preprint arXiv:1301.3584, jan 2013.

[4] G. Desjardins, K. Simonyan & R. Pascanu, "Natural Neural Networks," *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2015.

[5] J. Martens, *New Insights and Perspectives on the Natural Gradient Method*, arXiv preprint arXiv:1301.3584, Dec. 2014.

[6] T.N. Sainath, B. Kingsbury & H. Soltau, "Optimization Techniques to Improve Training Speed of Deep Neural Networks for Large Speech Tasks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2267–2276, 2013.

Residual Memory Networks: Feed-forward approach to learn long temporal dependencies

Murali Karthick Baskar, Martin Karafiát, Lukáš Burget

Brno University of Technology, Speech@FIT and IT4I COE, Brno, Czech Republic
{baskar,karafiat,burget}@fit.vutbr.cz

Abstract

Training deep recurrent neural network (RNN) architectures is complicated due to the increased network complexity. This disrupts the learning of higher order abstracts using deep RNN. In case of feed-forward networks training deep structures is simple and faster while learning long-term temporal information is not possible. In this work we propose a residual memory neural network (RMN) architecture to model short-time dependencies using deep feed-forward layers having residual and time delayed connections. The number of layers in RMN signifies both the hierarchical processing depth and temporal depth. The key points are:

- The use of residual connections after every few layers to increase the network depth makes training faster with increase in performance. During backpropagation, the residual lines allows unimpeded flow of gradients and the time delay units capture temporal information with shared weights.
- A memory component is present in each layer in a serial manner where the first layer sees $t-T$ time instant and the last layer sees $t-1$ frame. The component weights are shared across all layers to enable them to learn longer-context.

The computational complexity in training RMN is significantly less when compared to deep recurrent networks due to its feed-forward design. RMN is further extended as bi-directional RMN (BRMN) by adding an extra connection with shared weights for learning future information. Computational complexity is relatively less for BRMN over BLSTM or bi-RNNs.

Recognition performance of RMN trained with 300 hours of Switchboard corpus is compared with various state-of-the-art LVCSR systems. The results in table 1 indicate that RMN and BRMN gains 6 % and 3.8 % relative improvement over LSTM and BLSTM networks. Further investigation is done using the Babel corpus for Pashto Language. Here, the input is 80 dimensional stacked bottleneck (SBN)_1stage features and 100 dimensional ivectors. The effectiveness of the uni-directional RMN model for language modeling is also investigated in [1] and the perplexity scores are compared against LSTM and simple RNN.

Table 1: Comparison of RMN with existing methods in literature trained using 300 hours of Switchboard corpus and tested with Hub5-00 eval set. In this table 3g is trigram, 4g is meant as 4-gram, bn-fMLLR is bottleneck features with fMLLR and ivec represents 100 dimensional ivectors.

% WER	Model Type	SWB (% CE WER)		
		3g	4g	4g+ivec
Proposed Models	RMN	13.0	12.0	**10.9**
	BRMN	11.8	10.8	**9.9**
State-of-the-art results in literature	TDNN [2]			12.5
	Unfolded RNN + fMLLR [3]			12.7
	LSTM + bn-fMLLR [4]			10.8
	LSTM [2]			11.6
	BLSTM [2]			10.3

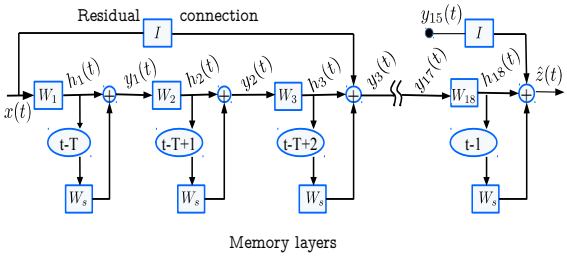


Figure1 :Architecture of residual memory network (RMN) with number of memory layers $L = 18$.

References

[1] Lukáš Burget Karel Beneš, Murali Baskar, "Residual memory networks in language modeling: Improving the reputation of feed-forward networks," in *INTERSPEECH*, 2017.

[2] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," *INTERSPEECH*, 2016.

[3] George Saon, Hagen Soltau, Ahmad Emami, and Michael Picheny, "Unfolded recurrent neural networks for speech recognition.," in *INTERSPEECH*, 2014, pp. 343–347.

[4] George Saon, Tom Sercu, Steven J. Rennie, and Hong-Kwang Jeff Kuo, "The IBM 2016 english conversational telephone speech recognition system," *CoRR*, vol. abs/1604.08242, 2016.

ASR teacher-student training and ensemble target diversity

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

University of Cambridge, United Kingdom

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

Abstract

Teacher-student training is a method to compress a large model or ensemble of models into a single student model, for computational efficiency during recognition. A standard method to train the student in Automatic Speech Recognition (ASR) is to minimise the KL-divergence between the teacher and student frame posteriors, represented by the outputs of the Neural Network (NN) acoustic models. However, this training method assumes that both the student and teacher use the same output target representations.

One scenario where different targets are required is when compressing an ensemble that uses a diversity of output targets. In ASR, significant performance gains can often be obtained by using ensemble methods. Ensemble methods can be viewed as a Monte Carlo approximation to Bayesian inference, by taking a combination over a finite number of models. The ensemble can approximate the uncertainty about the model parameters, captured within the diversity of the models used. When constructing an ensemble, it is important to consider how the models should be made different, to best leverage upon their diverse and complementary behaviours. Diversity can be introduced using methods such as multiple Random Initialisations (RI), Dropout, AdaBoost, and using different model architectures. This paper focuses on the method of using different output targets between models. In the standard ASR hybrid architecture, the NN acoustic models are usually trained to discriminate between different clusters of context dependent states at the outputs. A standard method of constructing these clusters of states is to use a Phonetic Decision Tree (PDT). A commonly used method to train the PDT is to select the greedy split at each node. This training procedure can be modified to produce multiple PDTs. One such modification is the Random Forest (RF) method, which uniformly samples from the n-best splits at each node. The ensemble can be constructed by associating a separate PDT with the output of each NN. Each NN then learns to discriminate between its own set of state clusters. It is hoped that such an ensemble will exhibit good diversity and complementarity between the behaviours of the models. The experiments suggest that the target diversity method is especially beneficial when the quantity of training data is limited and the PDTs are small.

Although the ensemble may perform well, it can be computationally expensive to use during recognition. Teacher-student training can reduce this computational demand. To allow the use of different output targets between the teachers and student, this paper proposes to train the student by minimising the KL-divergence between logical context posteriors, instead of frame posteriors [**?**]. This criterion is shown to reduce to the standard teacher-student criterion, but with the teachers' posteriors being mapped to the student's PDT. This formulation therefore allows teacher-student training to be used when the output targets of the teachers and student differ. With this method, it is also possible to select the output complexity of the student model independently of those of the teachers.

Table 1: Teacher-student training

Dataset	Ensemble method	Student WER (%) TS	+ sMBR	Ensemble WER (%)
Tok Pisin VLLP	RI	46.9	46.6	46.7
	RF	47.3	46.6	46.0
Javanese FLP	RI	52.4	51.6	52.5
	RF	52.7	51.9	52.4
HUB4	RI	8.9	8.8	8.8
	RF	9.2	9.0	8.7

The proposed method is evaluated on the HUB4 English broadcast news task, and both the Tok Pisin and Javanese conversational telephone speech tasks from the IARPA Babel programme. The very limited language pack is used for Tok Pisin, to provide a scenario with very limited training resources. The RF method is used to obtain multiple PDTs for the ensemble. Standard cross-entropy training of single models gives WERs of 50.2 %, 55.9 %, and 10.0 % for Tok Pisin, Javanese, and HUB4 respectively. The results in Table 1 show that by using the proposed criterion, the student performance can be brought closer to that of the RF teacher ensemble, than through standard cross-entropy training. However, the results also show a consistently larger performance degradation between the student and teacher ensemble when the PDTs differ. It is hypothesized that this degradation may either be due to the posterior mapping process or the limited output complexity of the student model.

Table 2: RF ensemble students with larger PDTs, for Tok Pisin

Student PDT size	Student WER (%) TS	+ sMBR	Ensemble WER (%)
1000	47.3	46.6	
1800	47.0	46.3	46.0
15094 (RF tied states)	46.6	46.0	

Both these issues can be addressed by using a student with a more complex output. Indeed, the experimental results in Table 2 show that using a larger PDT for the student brings its performance closer to that of the RF teacher ensemble. By using a larger student PDT, the student is better able to leverage upon the RF ensemble's good performance, and outperforms the student of the RI ensemble.

1. References

[1] J. H. M. Wong and M. J. F. Gales, "Student-teacher training with diverse decision tree ensembles," in *INTERSPEECH*, Stockholm, Sweden, Aug 2017.

Simplifying very deep convolutional neural network architectures for robust speech recognition

Joanna Rownicka, Steve Renals, Peter Bell

The Centre for Speech Technology Research, University of Edinburgh, UK

j.m.rownicka@sms.ed.ac.uk, {s.renals, peter.bell}@ed.ac.uk

Abstract

Very deep convolutional neural networks (VDCNNs) have been successfully used in computer vision. More recently VDCNNs have been applied to speech recognition, using architectures adopted from computer vision. In this paper, we experimentally analyse the role of the components in VDCNN architectures for robust speech recognition. We have proposed a number of simplified VDCNN architectures, taking into account the use of fully-connected layers and down-sampling approaches. We have investigated three ways to down-sample feature maps: max-pooling, average-pooling, and convolution with increased stride. Our proposed model consisting solely of convolutional (conv) layers, and without any fully-connected layers, achieves a lower word error rate (WER) on Aurora 4 compared to other VDCNN architectures typically used in speech recognition. Our model trained with alignments generated from synchronized clean-condition training set achieves WER = 7.75%, and the model trained with multi-condition targets achieves WER = 9.55%. The relative improvements over the baseline VDCNN models are 11% and 5% respectively. We have also extended our experiments to the MGB-3 task of multi-genre broadcast recognition using BBC TV recordings. The MGB-3 results indicate that the same architecture achieves the best result among our VDCNNs on this task as well.

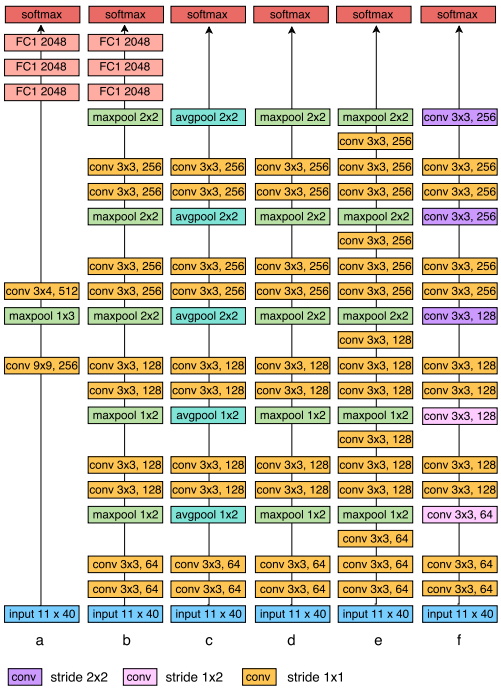


Figure 1: Architectures of the convolutional models: a) *CNN*, b) *VDCNN-max-4FC*, c) *VDCNN-avg*, d) *VDCNN-max*, e) *VDCNN-max-addconv*, f) *VDCNN-allconv*.

Future Word Contexts in Neural Network Language Models

Xie Chen

August 18, 2017

Abstract

Recently, bidirectional recurrent network language models (bi-RNNLMs) have been shown to outperform standard, unidirectional, recurrent neural network language models (uni-RNNLMs) on a range of speech recognition tasks. This indicates that future word context information beyond the word history can be useful.

However, bi-RNNLMs pose a number of challenges as they make use of the complete previous and future word context information. This impacts both training efficiency and their use within a lattice rescoring framework. In this paper these issues are addressed by proposing a novel neural network structure, succeeding word RNNLMs (su-RNNLMs). Instead of using a recurrent unit to capture the complete future word contexts, a feedforward unit is used to model a finite number of succeeding, future, words. This model can be trained much more efficiently than bi-RNNLMs and can also be used for lattice rescoring.

Experimental results on a meeting transcription task (AMI) show the proposed model consistently outperformed uni-RNNLMs and yield only a slight degradation compared to bi-RNNLMs in N-best rescoring. Additionally, Consistent and significant improvements (0.4-0.5% absolute) can be obtained using lattice rescoring and subsequent confusion network decoding over standard unidirectional RNNLMs.

Automatic speech recognition for media monitoring in SUMMA

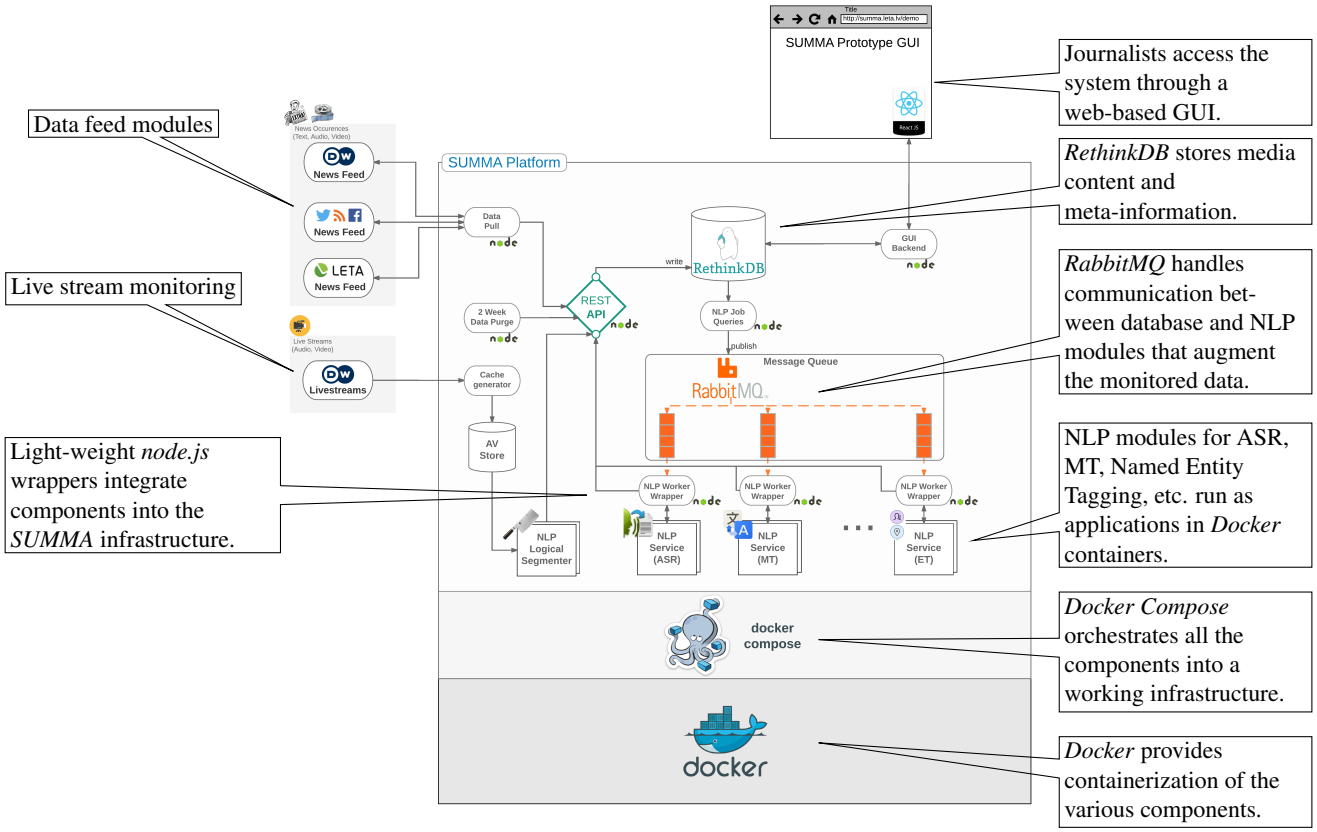
Peter Bell, Ondřej Klejch, Steve Renals

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

1. Abstract

We present a prototype of the SUMMA platform [1]: an integrated platform for multilingual media monitoring. The platform contains a rich suite of low-level and high-level natural language processing technologies, including automatic speech recognition of broadcast media, machine translation, automated tagging and classification of named entities, semantic parsing to detect relationships between entities, and automatic construction of factual knowledge bases.

This presentation will focus on the development of automatic speech recognition for broadcast media used in the SUMMA platform. Speech recognition operates in an online fashion using the *CloudASR* tool [2]. The platform will eventually support nine languages, both well-resourced (such as English and Arabic) and less well-resourced (including Farsi, Ukranian and Latvian). We will discuss issues such as lightly supervised training and cross-lingual adaptation.



2. References

[1] R. Liepins, U. Germann *et al.*, "The SUMMA platform prototype," in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 116–119. [Online]. Available: http://aclweb.org/anthology/E17-3029

[2] O. Klejch, O. Plátek, L. Žilka, and F. Jurčíček, "CloudASR: platform and service," in *Int'l. Conf. on Text, Speech, and Dialogue*. Springer, 2015, pp. 334–341.

Improving Automatic Speech Recognition for Spoken Language Assessment

Y. Wang, A. Ragni, X. Chen, K. M. Knill, M. J. F. Gales

Automatic spoken language assessment systems need fast and accurate automatic speech recognition (ASR) systems for deployment and feedback. However, it is challenging to obtain good ASR performance on non-native English learners speech because they have various first languages (L1), accents and proficiency levels. First, phonetic lexica, which are normally used in English ASR systems, reflect standard English pronunciations and may not be appropriate for this data. As an alternative, graphemic lexica may better model the non-native pronunciations. Second, it is challenging to transcribe such data due to low inter-annotator agreement and high expense of transcribing large volumes of available data, thus semi-supervised training and active learning could be leveraged. In this paper, we will explore those techniques and their impacts on improving the ASR systems.

Two examples will be given on how improved ASR systems will affect the downstream tasks, in which we compare two types of ASR systems. One is a baseline speaker-independent DNN hybrid system that is trained on 100 hours of Gujarati data (referred as the *basic* configuration). The other one is a more advanced speaker independent joint system (shown in Fig. 1) that is trained on 300 hours of mixed L1 data (referred as the *advanced* configuration). This advanced system utilised semi-supervised training, active learning, graphemic lexicon and advanced forms of acoustic models. The performance of the two configurations are listed in Tab. 1. We compare the tree kernel similarity scores calculated from the hypothesis generated by the advanced configuration (Hyp1) and those calculated from the hypothesis generated by the basic configuration (Hyp2). From Fig. 2 it can be seen that Hyp1 gives a consistent improvement over Hyp2 and performs similarly to the crowd-sourced transcription (CWD). For grading, we extract term frequency–inverse document frequency (TFIDF) based features from the Part-of-Speech (POS) tags generated by the two hypothesis. In Tab. 2, it can be seen that when the POS tags are generated from the advanced configuration and has a lower error rate, the TDIDF features can give an improved grader performance over the baseline features.

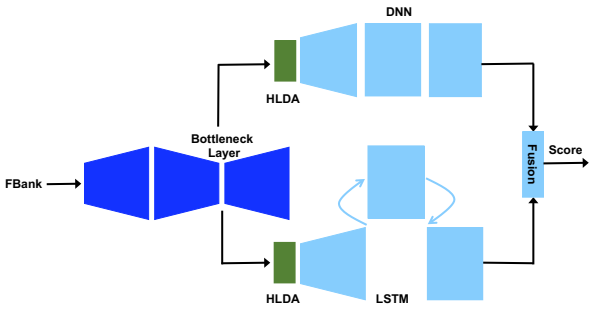


Figure 1: Joint system build in Kaldi.

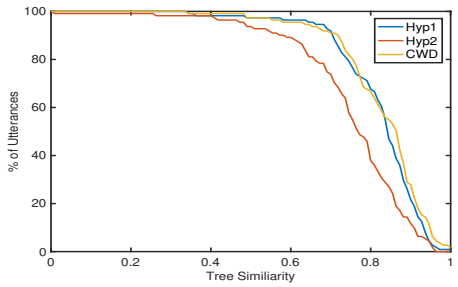


Figure 2: Tree kernel similarities comparing two ASR systems.

Configuration	%WER		
	eval1	eval2	eval3
Basic	36.4	50.9	47.5
Advanced	30.1	30.8	30.4

Table 1: ASR performance for two configurations.

Configuration	%WER	Features	PCC
Basic	47.5	Baseline	0.854
		+TFIDF	0.847
Advanced	30.4	Baseline	0.849
		+TFIDF	0.860

Table 2: Grader performance for two ASR configurations.

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.

A learned emotion space for emotion recognition and emotive speech synthesis

Zack Hodari, Simon King

The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK
zack.hodari@ed.ac.uk, Simon.King@ed.ac.uk

1 Abstract

Emotion is a complex phenomenon that contributes heavily to human communication. Typically, human-computer interaction and text-to-speech systems do not account for emotion information, possibly due to lack of accurate emotion recognition and emotive speech synthesis methods. It seems likely that emotion recognition and synthesis has the ability to greatly improve how humans interface with machines.

Existing methods in speech recognition make use of a wide range of features and models. However, these methods make use of either categorical, or appraisal-based emotion descriptions. We believe these have flaws that limit their ability to describe emotion.

We present several neural network models for emotion recognition. In addition, we propose an abstract emotion space that avoids the flaws of existing emotion descriptions. We use stimulation to improve the interpretability of our emotion space, along with multi-task learning to improve its robustness. Finally, we investigate auxiliary features for style adaptation in statistical parametric speech synthesis, evaluating both our emotion space and other descriptions of emotion.

The results indicate that our recognition models are state-of-the-art. However, evaluation using speech synthesis shows that our emotion space is no more informative than existing emotion descriptions. Additionally, we investigate a convolutional recognition model using the spectrogram, which outperforms other spectrogram based methods.

Fine Tuning Neural Network Based Phoneme Recognition System For Children's Speech Therapy

Revathy Nayar†, Dr Mark Dunlop†, Prof Anja Lowit‡, Prof John Soraghan¤, Dr John Levine†

† Department of Computer and Information Science
‡ School of Psychological Sciences and Health
¤ Department of Electronic and Electrical Engineering
University of Strathclyde, United Kingdom
revathy.nayar@strath.ac.uk

Abstract

Recent studies in the UK have shown that 6% of children have some form of speech and language difficulty [1].Whilst most children's difficulties resolve over time, children whose speech anomalies persist into primary school may have long-term problems concerning literacy, social interaction and school attainment[2] . Advances in speech technology can be used to develop assistive aids for child under therapy for speech sound disorders. Therapy sessions for children's speech sound disorder is conducted in face-to-face sessions with the speech language therapist(SLT) eliciting speech that includes the problematic segment that has been targeted for intervention. This problematic speech segments can be analysed automatically through phoneme recognition. Previous studies have attempted to develop assessment tools for speech disorders for children mostly to be used in a clinical setting only and cannot be used without expert supervision such as for home practise [3].

In this study, a phoneme recognition system that performs a phoneme recognition on English stimuli is being researched. Identification and extraction of the main acoustic features and thereby classification to indicate the accuracy levels of the speech sound produced can be used to give feedback on therapy progress. The proposed method involves a pre-processing stage where audio samples are normalised followed by a feature extraction stage. The features so extracted is then classified by a single layer neural network. The challenges relating to data sparsity and personalization are being addressed in this work which are the two main challenges of adopting speech technology for children's clinical applications[4]. Further user-centric interaction design, along with knowledge from speech therapy principles can complement speech processing in ways that compensate for the lack of 100% accuracy observed in such recognition systems[5]. This technology along with computer gaming techniques can be used in the future to develop engaging speech therapy tools for children.

References

[1] thecommunicationtrust.org, 'Communication Difficulties - facts and stats.pdf', April 2011.
[2] Y. Wren, L. L. Miller, T. J. Peters, A. Emond, and S. Roulstone, 'Prevalence and Predictors of Persistent Speech Sound Disorder at Eight Years Old: Findings From a Population Cohort Study', *J Speech Lang Hear Res*, vol. 59, no. 4, pp. 647–673, August 2016.
[3] N. R. Silton, *Recent Advances in Assistive Technologies to Support Children with Developmental Disorders*. IGI Global, 2015.
[4] D.-V. Popovici and C. Buică-Belciu, 'Professional challenges in computer-assisted speech therapy', *Procedia - Social and Behavioral Sciences*, vol. 33, pp. 518–522, 2012.
[5] C. Munteanu *et al.*, 'Designing Speech, Acoustic and Multimodal Interactions', 2017, pp. 601–608.

Automatic Analysis of Motivational Interviewing with Diabetes Patients

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

Abstract

Motivational Interview (MI) is one kind of goal-driven Clinical Conversion (CCs) between a clinician and patient that seeks to facilitate and engage a patient's intrinsic motivation to change his or her behavior. It's a significant part of the treatments of patients. Clinicians are trained in clinical conversations, which focus on diagnosis (history taking) and communication, but at present it is not possible to monitor their effectiveness in following good practice. There are clear established guidelines for how conversations should be carried out. Conventional assessment of MI is based on human analysis of the transcriptions of a clinician-patients dialogue to measure the clinician's compliance with guidelines, which is labor-intensive, time-consuming and financial costing. Natural language processing techniques have been utilized for modeling clinicians' empathy in motivational interview [1]. A Deep Learning Approach is used to predict the empathy rating from transcripts of the interviews in [2].

In this study, a DNN-HMM Automatic Speech Recognition (ASR) system was built to replace the manual transcriptions by automatic transcriptions. The system was trained on the AMI meeting corpus and 7 hours of recordings of MI sessions with diabetes patients. By further training the network using the MI data only, we achieved a 49.92% Word Error Rate, compared with 55.38% for the system trained only on AMI corpus and 53.13% for the baseline system trained on both AMI and MI data.

A speaker diarization system is also being built to distinguish the clinicians' speech and the patients' speech. The GMM I-vectors extracted from the manual segmentation showed a 94.8% accuracy on clustering, which motivated us to apply the I-vector approach to automatic segmentation.

Reference:

[1] B. Xiao, P. G. Georgiou, and S. Narayanan, " Analyzing the language of therapist empathy in motivational interview based psychotherapy," in Asia-Pacific Signal and Information Processing Association, 2012, pp. 1-4

[2] Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. (2016). A deep learning approach to modeling empathy in addiction counseling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-September-2016*, 1447-1451.

Name: PyDial: A Multi-domain Statistical Dialogue System Toolkit

Authors: Stefan Ultes, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrksic, Tsung-Hsien Wen, Milica Gasic and Steve Young

Affiliation: University of Cambridge

Abstract:

Designing speech interfaces to machines has been a focus of research for many years. These Spoken Dialogue Systems (SDSs) are typically based on a modular architecture consisting of input processing modules speech recognition and semantic decoding, dialogue management modules belief tracking and policy, and output processing modules language generation and speech synthesis (see Fig. 1).

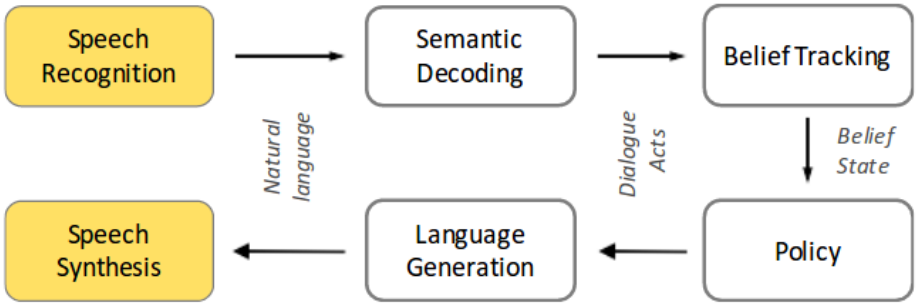


Figure 1: Architecture of a modular SDS

Statistical SDS are speech interfaces where all SDS modules are based on statistical models learned from data [1][2] (in contrast to hand-crafted rules). Despite the rich body of research on statistical SDS, there is still no common platform or open toolkit available. The availability of a toolkit targetted specifically at statistical dialogue systems would enable people new to the field would be able to get involved more easily, results to be compared more easily, and researchers to focus on their specific research questions instead of reimplementing algorithms (e.g., evaluating understanding or generation components in an interaction).
Hence, to stimulate research and make it easy for people to get involved in statistical spoken dialogue systems, we present PyDial, a multi-domain statistical spoken dialogue system toolkit. PyDial is implemented in Python and is actively used by the Cambridge Dialogue Systems Group.
PyDial supports multi-domain applications in which a conversation may range over a number of different topics. This introduces a variety of new research issues including generalised belief tracking, rapid policy adaptation and parallel learning, and natural language generation.

[1] Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. Proceedings of the IEEE 101(5):1160–1179.

[2] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In Proceedings of EACL

From media fears to research reality: how ready are countermeasures against speaker verification spoofing?

Jasmin Liu[1], *Federico Alegre*[1], *Benoit Fauve*[1], *Anssi Kanervisto*[2], *Md Sahidullah*[2], *Tomi Kinnunen*[2], *Héctor Delgado*[3], *Massimiliano Todisco*[3], *Nicholas Evans*[3], *Mauro Falcone*[4]

[1]ValidSoft, UK
[2]University of Eastern Finland, Finland
[3]EURECOM, France
[4]Fondazione Ugo Bordoni, Italy

{jasmin.liu, federico.alegre, benoit.fauve}@validsoft.com
anssi.kanervisto@uef.fi, {sahid, tkinnu}@cs.uef.fi,
{delgado, todisco, evans}@eurecom.fr, falcone@fub.it

Abstract

The subject of biometrics spoofing has received more attention over the last few years within the research community. For speech technology spoofing attacks [1] relate to replay attacks [2, 6] and a category sometimes referred to as synthetic voice attacks, including attacks with artificial signals such as voice conversion and speech synthesis [3, 4, 5]. Some significant initiatives came with the ASVspoof2015 evaluation focused on spoofing with synthetic voice [3] and more recently ASVspoof2017, an evaluation focusing on replay detection [6].

We present some results about the latest development on voice anti-spoofing, most of them derived from work undertaken during the H2020 OCTAVE project [7]. Results on synthetic voice detection [Figure 2] are presented with synthetic audio from systems recently mentioned in press coverage [Figure 1]. More results are presented about the recent development of replay detection systems within the Octave project and for the ASVspoof2017 challenge. Finally, we discuss the challenges ahead, with results showing the difficulty of building up representative databases and further results showing adverse effects of codec and noise.

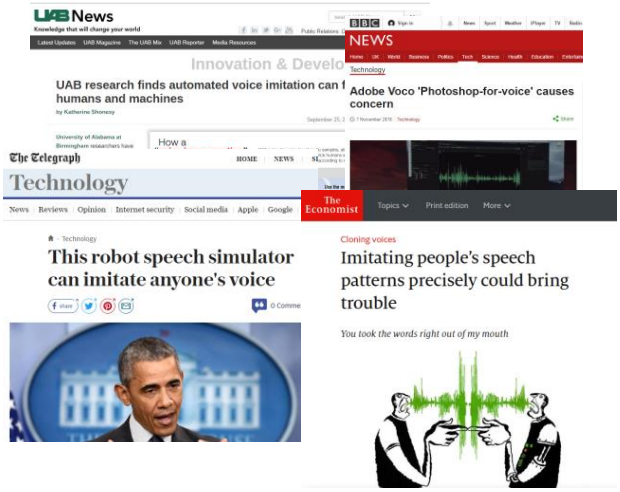


Fig. 1: Some of the recent coverage about vulnerability of voice biometrics systems from mainstream media including BBC, the Telegraph and the Economist.

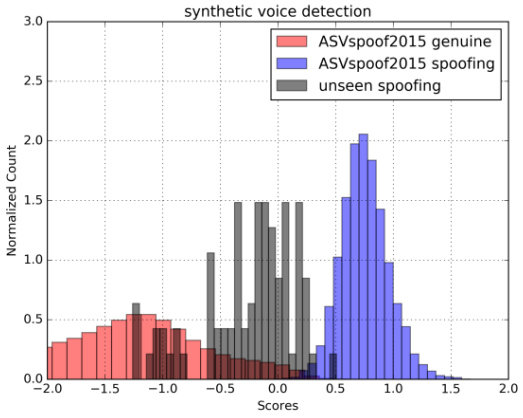


Fig. 2: Score distribution from a state of the art synthetic voice detection system on ASVSpoof2015 standard database (red and blue distributions) and on a set of 80 files (unseen set in grey distribution) from companies mentioned in recent press coverage.

References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, no. 0, pp. 130–153, 2015.

[2] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in BIOSIG 2014 - Proceedings of the 13th International Conference of the Biometrics Special Interest Group, 10.-12. September 2014, Darmstadt, Germany, 2014, pp. 157–168.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, "ASVspoof 2015: "The First Automatic Speaker Verification Spoofing and Countermeasures Challenge," in INTERSPEECH 2015, September 6-1, Dresden, Germany, 2015.

[4] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A Comparison of Features for Synthetic Speech Detection," in INTERSPEECH 2015, Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, 2015.

[5] M. Todisco, H. Delgado, and N. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," in ODYSSEY 2016, The Speaker and Language Recognition Workshop, June 21-24, Bilbao, Spain, 2016.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in INTERSPEECH 2017, August 20-24, 2017, Stockholm, Sweden, 2017.

[7] List of publications from the H2020 Octave project: https://www.octave-project.eu/documents/publications/

Evaluation of Speech Synthesis for Virtual Characters

João P. Cabral[1]*, Benjamin R. Cowan*[2]*, Katja Zibrek*[1]*, Rachel McDonnell*[1]

[1]Trinity College Dublin, Ireland
[2]University College Dublin, Ireland
cabralj@tcd.ie, benjamin.cowan@ucd.ie, zibrekka@tcd.ie, ramcdonn@tcd.ie

1. Abstract

Graphical realism and the naturalness of the voice used are important aspects to consider when designing a virtual agent or character. In this work, we evaluate how synthetic speech impacts people's perceptions of a rendered virtual character. Using a controlled experiment, we focus on the role that speech, in particular voice expressiveness in the form of personality, has on the assessment of voice level and character level perceptions. In the experiment, we used eight video clips that varied in the personality portrayed by an actor. To create the virtual character a 3D scan of the actor's face was taken using a structured light 3D scanner, as in [1]. We varied the voice the character was given (either human recording or synthetic voice), keeping the graphical rendering method for the character constant in all conditions. The synthetic speech was produced by copy-synthesis using the Uniform Concatenative Excitation Model [2]. Participants were asked to watch the videos and after each video they were asked to rate aspects of the character in general and aspects specific to the character's voice. We found that people rated a real human voice as more expressive, understandable and likeable than the expressive synthetic voice we developed. Contrary to our expectations, we found that the voices did not have a significant impact on the character level judgments; people in the voice conditions did not significantly vary on their ratings of appeal, credibility, human-likeness and voice matching the character. The implications this has for character design and how this compares with previous work are discussed.

2. Acknowledgements

This research is supported by Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie) at Trinity College Dublin, the UCD Seed Funding Scheme (Grant SF1191) and the Irish Research Council funded Cog-SIS project (Grant R17339).

3. References

[1] R. McDonnell, M. Breidt, and H. H. Bülthoff, "Render Me Real?: Investigating the Effect of Render Style on the Perception of Animated Virtual Humans," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 91:1–91:11, 2012.

[2] J. P. Cabral, "Uniform Concatenative Excitation Model for Synthesising Speech without Voiced/Unvoiced Classification," in *14th Annual Conference of the International Speech Communication Association INTERSPEECH*, Lyon, France, pp. 1082–1085, 2013.

A Hierarchical Encoder-Decoder Model for SPSS

Srikanth Ronanki, Oliver Watts, Simon King

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

srikanth.ronanki@ed.ac.uk

1. Abstract

Current approaches to statistical parametric speech synthesis using Neural Networks generally require input at the same temporal resolution as the output, typically a frame every 5ms, or in some cases at waveform sampling rate. It is therefore necessary to fabricate highly-redundant frame-level (or sample-level) linguistic features at the input. This paper proposes the use of a hierarchical encoder-decoder model to perform the sequence-to-sequence regression in a way that takes the input linguistic features at their original timescales, and preserves the relationships between words, syllables and phones. The proposed model is designed to make more effective use of supra-segmental features than conventional architectures, as well as being computationally efficient. Experiments were conducted on prosodically-varied audiobook material because the use of supra-segmental features is thought to be particularly important in this case. Both objective measures and results from subjective listening tests, which asked listeners to focus on prosody, show that the proposed method performs significantly better than a conventional architecture that requires the linguistic input to be at the acoustic frame rate.

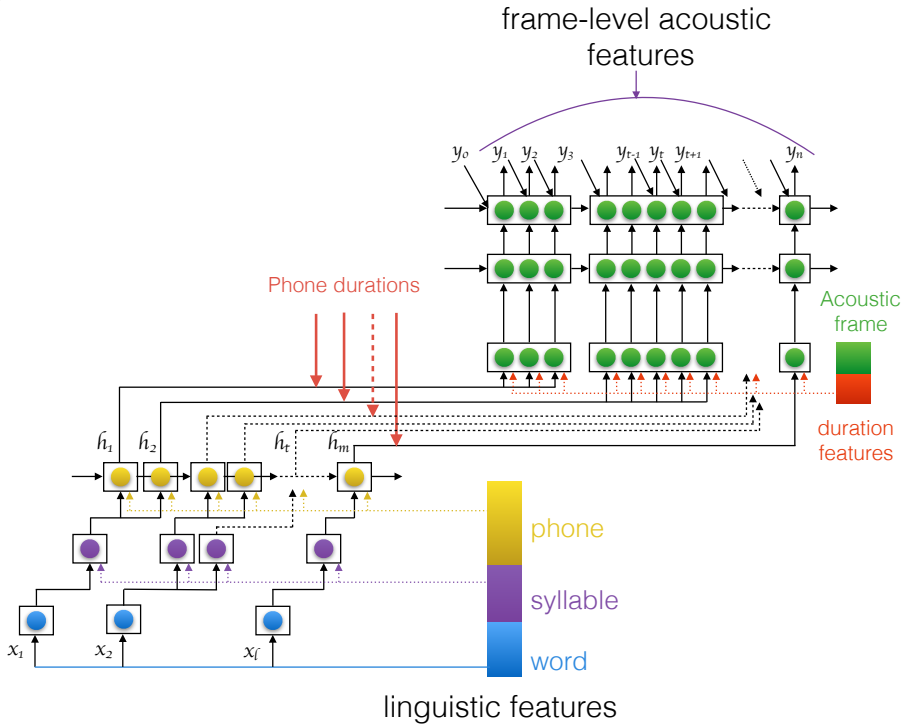


Figure 1: Schematic diagram of a hierarchical encoder-decoder for SPSS. The lower part of the network is the hierarchical encoder, with each layer operating at a particular linguistic level, and phone-level recurrence as its final encoded output. The upper part of the network is the decoder, generating speech parameters using frame-level recurrence. Solid black lines indicate the propagation of hidden activations between layers, and dashed colored lines indicate the injection of linguistic features at the appropriate level. The patterns of connections between word, syllable and phone layers is determined by the known linguistic structure of the current utterance. Each block of green units represents a phone, with the number of units corresponding to its duration in frames (although not drawn to scale).

We combine a hierarchical *encoder* to model linguistic structure at multiple time-scales, with a *decoder* that predicts speech parameters for each output acoustic frame making use of sequence transduction networks [1].

2. References

[1] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711 (not peer reviewed)*, 2012.

Integrated speaker-adaptive speech synthesis

Moquan Wan, Gilles Degottex, Mark J.F. Gales

Cambridge University Engineering Department, UK

mw545@cam.ac.uk, gad27@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Enabling speech synthesis systems to rapidly adapt to sound like a particular speaker is an essential attribute for building personalised systems. For deep-learning based approaches, this is difficult as these networks use a highly distributed representation. It is not simple to interpret the model parameters, which complicates the adaptation process. To address this problem, speaker characteristics can be encapsulated in fixed-length speaker-specific Identity Vectors (iVectors), which are appended to the input of the synthesis network. Altering the iVector changes the nature of the synthesised speech. The challenge is to derive an optimal iVector for each speaker that encodes all the speaker attributes required for the synthesis system. The standard approach involves two separate stages: estimation of the iVectors for the training data; and training the synthesis network. This paper proposes an integrated training scheme for speaker adaptive speech synthesis. For the iVector extraction, an attention based mechanism, which is a function of the context labels, is used to combine the data from the target speaker. This attention mechanism, as well as nature of the features being merged, are optimised at the same time as the synthesis network parameters. This should yield an iVector-like speaker representation that is optimal for use with the synthesis system. The system is evaluated on the Voice Bank corpus. The resulting system automatically provides a sensible attention sequence and shows improved performance from the standard approach.

Visual speech recognition: aligning terminologies

Helen L Bear and Sarah L Taylor
University of East London, London E16 2RD. University of East Anglia, Norwich NR4 7TJ

We are at an exciting time for machine lipreading. Traditional research stemmed from the adaptation of audio recognition systems. But now, the computer vision community is also participating. This joining of two previously disparate areas with different perspectives on computer lipreading is creating opportunities for collaborations, but in doing so the literature is sometimes conflicted with multiple uses of some terms and phrases.

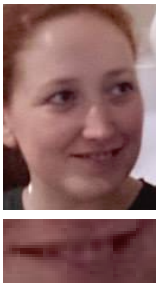


Figure 1. ROIs

Despite commonly being used interchangeably, the terms **speechreading** and **lipreading** have subtle but distinctive definitions. Speechreading is what human lip readers do. They interpret speech using information provided by the whole face and body since knowledge of the facial expression, gaze and body gestures can help to provides semantic context that makes decoding the speech easier. In the computer science, machine speechreading systems use the face (Fig 1 (top)). Lipreading is the interpretation of speech from the motion of the lips alone (Fig 1 (bottom)). This is the region of the image does not contain any information regarding the upper facial expression or body language.

Speaker independence in machine lipreading is achieved when classification models generalise to spoken utterances by talkers not contained within the training set (Fig 2). If a system only works on a closed set of speakers, or is not tested on speakers that are outside of the training set, we can assume that the approach is speaker dependent. For classification methods that require the data to be divided into train, validation, and test sets, the validation set can contain speakers from the training set and new speakers, but speakers must remain distinct from the test set if speaker independence is the goal. See Fig 3.

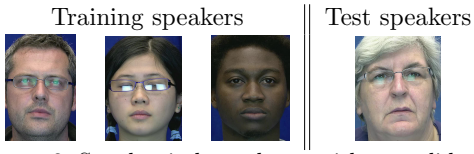


Figure 2. Speaker independence without validation

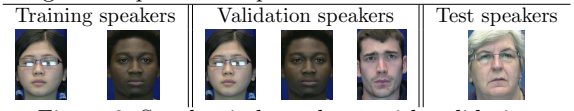


Figure 3. Speaker independence with validation

Methods of **scoring** the performance of machine lipreading have been adopted from audio speech recognition systems. The most common published figures are correctness, $C = \frac{N-D-S}{N}$ and accuracy, $A = \frac{N-D-S-I}{N}$. Previous literature reports lipreading performance based on different units (words, phonemes, or visemes) which makes comparing performance difficult. We are not in a position to definitively select the 'right' units, nor dictate the 'best' metric to use as that is a choice for each researcher to decide. We can however make a simple recommendation which would help us to quickly and easily compare results. M_{cu} || M^{nu} where M is the metric (e.g. A, C, or ER), and subscript notation for a classifier output, or a superscript notation for a network/dictionary output.

It is not uncommon in computer vision literature to report top five values as measure of classification performance. However, the inter-class variation in a set of phonemes or visemes is much smaller than, for example, that within a set of images representing, say cats, dogs, building, or landscape. If a cat and a dog image are confused, this is more reasonable than a confusion between a cat and a building. However, the issue this causes in speech, is that the classified output transcript can have a significantly different meaning resulting in confusion between talkers. Imagine if one was saying 'Pass me the salt' but due to misclassification in the top five classes, we transcribed 'Pass me the malt'. It is only one phoneme different (a /m/ instead of an /s/) but this single substitution is significant. Thus we recommend that if the top five values are to be reported, then the top one value should be included also.

We **recommend** authors be explicit about whether they are developing speech or lipreading systems so that the community can effectively compare methodologies. We have clarified the definition of speaker dependent machine lipreading, and authors should carefully consider the split of training, validation and test data prior to model training. To compare performance we have suggested a simple new notation. We recommend that if the top five values are to be reported, then the top one accuracy should be included separately, alongside the unit {v, p, or w}.

As a **final thought**; the fundamental motivation for lipreading is the ability to understand speech when the audio channel is hampered by noise. Therefore is is essential that future work includes acoustically challenging speech environments where the audio channel can not be recognised.

Poster Session 3
Tuesday 12th September, 11:00-12:15, LR3

1. E. Tsunoo, P. Bell, S. Renals: "Hierarchical recurrent neural network for story segmentation"
2. Q. Li, C. Zhang, F. L. Kreyssig, P. C. Woodland: "Experimental studies on teacher-student training of deep neural network acoustic models"
3. L. Bai, P. Jančovič, M. Russell, P. Weber, S. Houghton: "Phone classification using a non-Linear manifold with broad phone class dependent DNNs"
4. C. Wu, M. J. F. Gales: "Deep activation mixture model for speech recognition"
5. E. Loweimi, J. Barker, O. S. Torralba, T. Hain: "Robust source-filter separation of speech signal in the phase domain"
6. K. Kyriakopoulos, K. M. Knill, M. J. F. Gales: "A hierarchical architecture for automatic pronunciation assessment of spontaneous non-native English speech based on phone distances"
7. A. Malinin, A. Ragni, K. M. Knill, M. J. F. Gales: "Incorporating uncertainty into deep learning for spoken language assessment"
8. E. Gilmartin, M. O'Reilly, C. Saam, B. R. Cowan, C. Vogel, N. Campbell: "Silence and overlap in multiparty casual conversation"
9. D. Websdale, B. Milner: "Using visual speech information for speech enhancement"
10. B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, H. Christensen: "An avatar-based system for identifying individuals likely to develop dementia"
11. M. Roddy, N. Harte: "Towards predicting dialog acts from previous speakers' non-verbal cues"
12. B. Chettri, B. L. Sturm: "Combining information from multiple sources for ASV-antispoofing"
13. F. Espic, C. Valentini-Botinhao, S. King: "MagPhase vocoder: Magnitude and phase analysis/synthesis for statistical parametric speech synthesis"
14. V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, T. Drugman: "Phrase break prediction for long-form reading TTS: exploiting text structure information"
15. C. Valentini-Botinhao, J. Yamagishi: "Speech intelligibility in cars: the effect of speaking style, noise and listener age"
16. G. Sterpu, N. Harte: "Towards lipreading sentences with active appearance models"

Hierarchical Recurrent Neural Network for Story Segmentation

Emiru Tsunoo[1,2]*, Peter Bell*[1]*, Steve Renals*[1]

[1]The University of Edinburgh, United Kingdom
[2]Sony Corporation, Japan

Emiru.Tsunoo@jp.sony.com, Peter.Bell@ed.ac.uk, S.Renals@ed.ac.uk

Abstract

A broadcast news stream consists of a number of stories and each story consists of several sentences. We capture this structure using a hierarchical model based on a word-level Recurrent Neural Network (RNN) sentence modeling layer and a sentence-level bidirectional Long Short-Term Memory (LSTM) topic modeling layer. First, the word-level RNN layer extracts a vector embedding the sentence information from the given transcribed lexical tokens of each sentence. These sentence embedding vectors are fed into a bidirectional LSTM that models the sentence and topic transitions. A topic posterior for each sentence is estimated discriminatively and a Hidden Markov model (HMM) follows to decode the story sequence and identify story boundaries. Experiments on the topic detection and tracking (TDT2) task indicate that the hierarchical RNN topic modeling achieves the best story segmentation performance with a higher F1-measure compared to conventional state-of-the-art methods. We also compare variations of our model to infer the optimal structure for the story segmentation task.

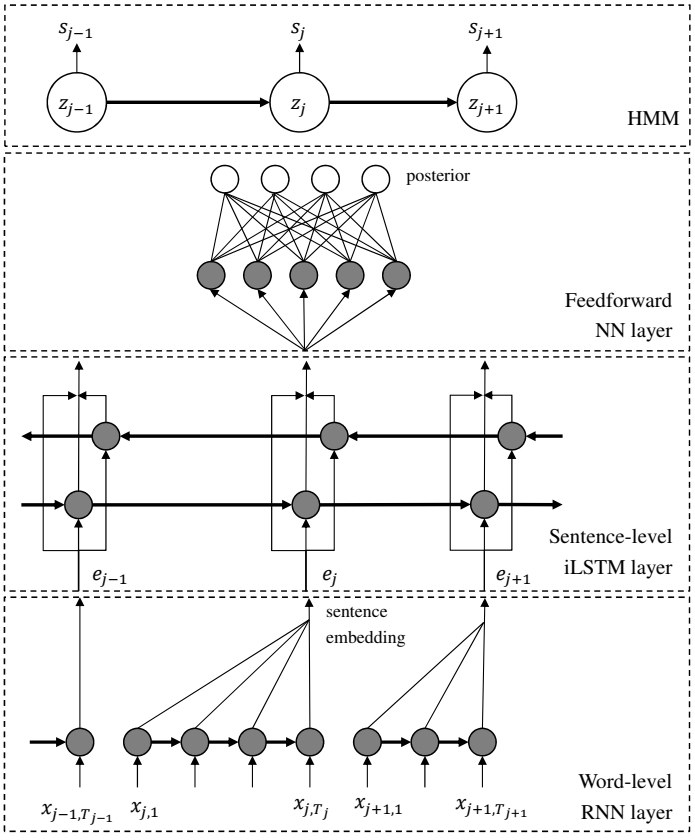


Figure 1: *Hierarchical recurrent neural network for story segmentation.*

Experimental Studies on Teacher-student Training of Deep Neural Network Acoustic Models

Q. Li, C. Zhang, F. L. Kreyssig and P. C. Woodland

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.
{ql264,cz277,flk24,pcw}@eng.cam.ac.uk

Abstract

Recently, Deep Neural Networks (DNNs) have been widely applied on various Automatic Speech Recognition (ASR) tasks, and have achieved state-of-the-art performance for acoustic modelling. However, as DNN models become more powerful, they also tend to be more complex in both width and depth, thus more computationally demanding. Therefore, training a small neural network with little compromise on performance is desirable. To this end, instead of training a small shallow model (the student) directly from hard classification labels as used in standard supervised learning, richer information obtained from soft labels generated by a large complex model (the teacher) is utilised [1] [2] [3]. By minimising the Kullback-Leibler divergence between the teacher and the student posterior distributions, it is demonstrated that teacher-student training approach yields better performance than learning directly from hard labels.

Table 1 shows that for a certain teacher model, the more complex the student model is, the better the student can mimic the knowledge of the teacher. Similarly, with identical architecture, the student model shows consistent gain when it learns from better teacher models (Table 2). It is further illustrated, in Table 3, that teacher-student training works for other types of deep neural networks, such as Recurrent Neural Networks (RNNs) [4]. Ensemble of complex models [5], which acts as the teacher, is able to guide a better student than each individual complex model could do. All experiments are conducted on the TIMIT dataset using the HTK toolkit, with 13-dimensional MFCC plus its first and second order differentials as input features. All models are evaluated on the full test set.

Student Arch.	T-S PER (%)	Baseline PER (%)
3-layer (250)	25.22	27.22
3-layer (500)	24.55	25.76
4-layer (500)	24.48	24.79

Table 1: Different fully-connected student architectures that learn from a 7-layer (500) fully-connected network with reference PER 23.55%.

Teacher Arch.	T-S PER (%)	Ref. PER (%)
7-layer (500)	24.55	23.55
7-layer (1000)	24.19	22.99
7-layer (1500)	24.01	22.49

Table 2: With baseline PER 25.76%, a 3-layer (500) fully-connected student model that learns from 7-layer fully-connected teacher models with increasing width.

Teacher Arch.	T-S PER (%)	Ref. PER (%)
7-layer (500)	24.55	23.55
RNN	23.84	20.59
Ensemble	**23.73**	**20.34**

Table 3: RNN (1 recurrent layer followed by a hidden and an output layer) and ensemble (linear ensemble between the 7-layer (500) fully-connected model and the RNN) teacher models, from which a 3-layer (500) fully-connected student model learns, with the student baseline PER being 25.76%.

References

[1] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, 2006, pp. 535–541.

[2] J. Ba and R. Caurana, "Do deep nets really need to be deep?" in *NIPS*, 2014, pp. 2654–2662.

[3] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *INTERSPEECH*, 2014, pp. 1910–1914.

[4] A. Graves, A. rahman Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.

[5] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *INTERSPEECH*, 2014, pp. 1915–1919.

Phone Classification using a Non-Linear Manifold with Broad Phone Class Dependent DNNs

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

Department of Electronic Electrical and Systems Engineering,
The University of Birmingham, Birmingham B15 2TT, UK

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

1. Abstract

Most state-of-the-art automatic speech recognition (ASR) systems use a single deep neural network (DNN) to map the acoustic space to the decision space. However, different phonetic classes employ different production mechanisms and are best described by different types of features. Hence it may be advantageous to replace this single DNN with several phone class dependent DNNs. The appropriate mathematical formalism for this is a manifold. This work extends our previous study of very low-dimensional bottleneck features (BNFs), including phone classification and BNF visualisation and interpretation [1, 2]. We assess the use of a non-linear manifold structure with multiple DNNs for phone classification. The system has two levels. The first comprises a set of broad phone class (BPC) dependent DNN-based mappings and the second level is a fusion network. Various ways of designing and training the networks in both levels are assessed, including varying the size of hidden layers, the use of the bottleneck or softmax outputs as input to the fusion network, and the use of different broad class definitions. Phone classification experiments are performed on TIMIT. The results show that using the BPC-dependent DNNs provides small but significant improvements in phone classification accuracy relative to a single global DNN. The paper concludes with visualisations of the structures learned by the local and global DNNs and discussion of their interpretations [3].

2. References

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

[2] P. Weber, L. Bai, M. Russell, P. Jančovič, and S. M. Houghton, "Interpretation of low dimensional neural network bottleneck features in terms of human perception and production," in *Interspeech*, San Francisco, CA, USA, 2016, pp.3384-3388

[3] L. Bai, P. Jančovič, M. Russell, P. Weber, and S. M. Houghton, "Phone Classification using a Non-Linear Manifold with Broad Phone Class Dependent DNNs," in *Interspeech*, Stockholm, Sweden, 2017.

Deep Activation Mixture Model for Speech Recognition

Chunyang Wu and Mark J.F. Gales

Cambridge University Engineering Dept
Trumpington St., Cambridge, CB2 1PZ, U.K.
{cw564, mjfg}@eng.cam.ac.uk

Abstract

Deep learning approaches achieve state-of-the-art performance in a range of applications, including speech recognition. However, the parameters of the deep neural network (DNN) are hard to interpret, which makes regularisation and adaptation to speaker or acoustic conditions challenging. This paper proposes the deep activation mixture model (DAMM) to address these problems. The output of one hidden layer is modelled as the sum of a mixture and residual models. The mixture model forms an activation function contour while the residual one models fluctuations around the contour. The use of the mixture model gives two advantages: first, it introduces a novel regularisation on the DNN; second, it allows novel adaptation schemes. The proposed approach is evaluated on a large-vocabulary U.S. English broadcast news task. It yields a slightly better performance than the DNN baselines, and on the utterance-level unsupervised adaptation, the adapted DAMM acquires further performance gains.

Acknowledgement

The research leading to these results was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012; research projects and donations from Google and Amazon. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the

ROBUST SOURCE-FILTER SEPARATION OF SPEECH SIGNAL IN THE PHASE DOMAIN

Erfan Loweimi, Jon Barker, Oscar Saz Torralba and Thomas Hain

Speech and Hearing Research Group (SPandH), University of Sheffield, Sheffield, UK

{{eloweimi1, j.p.barker, o.saztorralba,t.hain}@sheffield.ac.uk

ABSTRACT

Deconvolution of the excitation (source) and vocal tract (filter) components of the speech signal through cepstral processing is well-established and widely used in speech processing. In [1], we presented a novel source-filter decomposition via Trend-Fluctuation analysis of the phase spectrum of the minimum-phase part of speech, computed through Hilbert transform. It was demonstrated that the phase Trend and Fluctuation components have a clear relation to the vocal tract and excitation elements, respectively.

In the later work [2], the phase spectrum and its representations were studied through evaluating their statistical properties at various points along the parametrisation pipeline. It is demonstrated that speech phase spectrum has a bell-shaped distribution which is in contrast to the uniform assumption that is usually made. It was argued that the uniform density (which implies that the corresponding sequence is least-informative) is an artefact of the phase wrapping and is not an original characteristic of this spectrum.

Recently, in [3], further theoretical investigation and optimisations were proposed to make the source and filter representations less sensitive to the noise and better match to the downstream processing. To this end, first, in computing the Hilbert transform, the log function is replaced by the generalised logarithmic function. This introduces a tuning parameter that adjusts both the dynamic range and distribution of the phase-based representation. Second, when computing the group delay, a more robust estimate for the derivative is formed by applying a regression filter instead of using sample differences. The effectiveness of these modifications is evaluated in clean and noisy conditions by considering the accuracy of the fundamental frequency extracted from the estimated source, and the performance of features extracted from the estimated filter in ASR. In particular, the proposed filter-based front-end reduces Aurora-2 WERs by 6.3% (average 0-20 dB) compared with previously reported results. Furthermore, when tested in a CSR task (Aurora-4) the new features resulted in 5.8% absolute WER reduction compared to MFCCs without performance loss in the clean/matched condition.

Table 1. *WER for Aurora-4 (BN: BottleNeck) [3].*

Feature	A	B	C	D	*Ave*
MFCC [Clean]	7.4	34.5	29.4	50.3	39.0
Proposed [Clean]	7.1	26.5	28.1	45.1	33.2
BN-MFCC [Multi]	7.2	12.9	14.3	27.0	18.7
BN-Proposed [Multi]	7.0	12.7	14.3	26.6	18.4

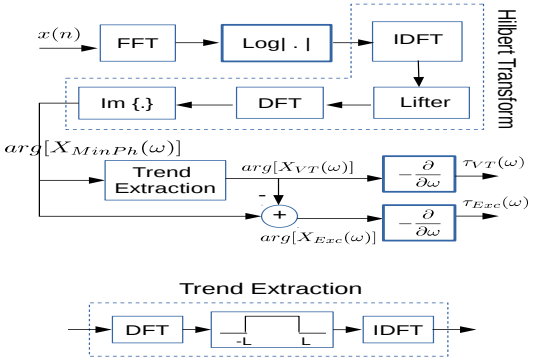


Fig. 1. Phase-based source-filter decomposition [1].

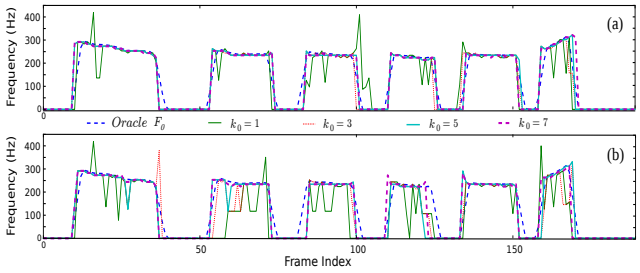


Fig. 2. Effect of k_0 (context length in regression filter) on the accuracy/robustness of phase-based F_0 estimation using SRH ($\alpha = 0.1$). (a) clean speech, (b) noisy (Gaussian white, 5 dB) speech [3].

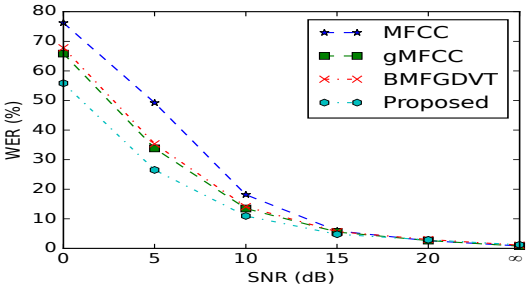


Fig. 3. Performance of different features vs SNR for Aurora-2 task (averaged over A, B and C testsets) [3].

1. REFERENCES

[1] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain.," in *Interspeech*, 2015.

[2] E. Loweimi, J. Barker, and T. Hain, "Statistical normalisation of phase-based feature representation for robust speech recognition," in *ICASSP*, March 2017, pp. 5310–5314.

[3] E. Loweimi, J. Barker, O. Saz Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," in *Interspeech*, 2017.

A hierarchical architecture for automatic pronunciation assessment of spontaneous non-native English speech based on phone distances

K. Kyriakopoulos, K.M. Knill, M.J.F. Gales

1. Abstract

Automated language proficiency assessment systems based on spontaneous non-native speech are becoming increasingly necessary to meet the demand for computer assisted language learning. In characterising pronunciation in particular, distances between phones have been shown to be indicative of speaker ability. These methods require large amounts of audio from each speaker, however, to obtain models for the manner of pronunciation of each phone.

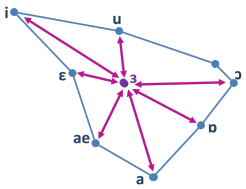


Figure 1: Illustration of the phone distance concept

This paper presents an approach for learning phone distance features for grading directly from the frame-level spectral (PLP) features of speaker audio using a hierarchical system based on Siamese networks. Audio of speakers' recorded responses to a spoken English test is passed through an automatic speech recogniser and time-aligned to a sequence of phones. The sequence of PLP features for each phone is projected to a feature vector using an RNN and an attention mechanism is used to combine feature vectors corresponding to different instances of the same phone. Euclidean distances between each possible pair of the resultant phone feature vectors are then projected to predict score, with the whole network trained through backpropagation. The performance of the network is compared to that of a DNN trained on baseline features and on phone distance features obtained from HMM models of each phone.

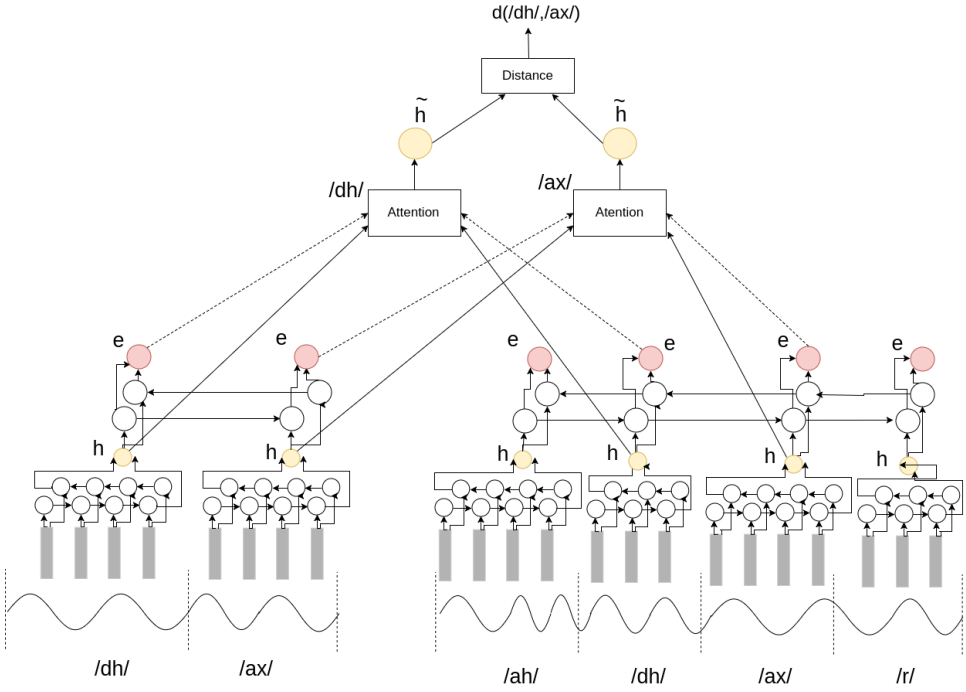


Figure 2: Illustration of deep phone distance architecture

Incorporating Uncertainty into Deep Learning for Spoken Language Assessment

Andrey Malinin, Anton Ragni, Kate M. Knill, Mark J. F. Gales
Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{am969, ar527, kate.knill, mjfg}@eng.cam.ac.uk

August 18, 2017

Abstract

There is a growing demand for automatic assessment of spoken English proficiency. These systems need to handle large variations in input data owing to the wide range of candidate skill levels and L1s, and errors from ASR. Some candidates will be a poor match to the training data set, undermining the validity of the predicted grade. For high stakes tests it is essential for such systems not only to grade well, but also to provide a measure of their uncertainty in their predictions, enabling rejection to human graders. Previous work examined Gaussian Process (GP) graders which, though successful, do not scale well with large data sets. Deep Neural Networks (DNN) may also be used to provide uncertainty using Monte-Carlo Dropout (MCD). This work proposes a novel method to yield uncertainty and compares it to GPs and DNNs with MCD. The proposed approach *explicitly* teaches a DNN to have low uncertainty on training data and high uncertainty on generated artificial data. On experiments conducted on data from the Business Language Testing Service (BULATS), the proposed approach is found to outperform GPs and DNNs with MCD in uncertainty-based rejection whilst achieving comparable grading performance.

Silence and overlap in multiparty casual conversation

Emer Gilmartin[1], *Maria O'Reilly*[2], *Christian Saam*[3], *Benjamin R. Cowan*[4], *Carl Vogel*[3], *Nick Campbell*[1]

[1]Speech Communication Lab, Trinity College Dublin, Ireland
[2]Trinity College Dublin, Ireland
[3]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[4]University College Dublin, Ireland

gilmare@tcd.ie, moreil12@tcd.ie, christian.saam@adaptcentre.ie, benjamin.cowan@ucd.ie, vogel@tcd.ie, vogel@tcd.ie

Abstract

Casual conversation, 'talk for the sake of talking', is a basic form of human spoken interaction, aiding social bonding. Understanding the structure of such talk is paramount for successful artificial casual or social spoken dialogue. Much social talk is multiparty and has no clear practical goal, with conversations lasting up to several hours. However, much available conversational data is task-based or of short duration. High levels of crosstalk and poor recording conditions make automatic analysis of natural multiparty conversation difficult. Casual conversation has been found to differ from written language and from more task-based interaction in terms of structure and content, and it is also likely that even basic features such as mechanisms of turntaking and prosodic patterning could vary. We describe our explorations into silence and overlap in casual conversation, using manually segmented long (c. 1 hr) informal multiparty conversations. We analyse the speech and silence activity following intervals where one participant speaks in the clear for a second or more, and categorise patterns of overlap and turn change or retention. We also report on a pilot study of a subset of our dataset comprising over 200 manually annotated intonational phrases (IP) adjacent to silences and overlaps, analysing IP-final tunes with the IViE intonational transcription system, and measuring IP duration to investigate prosodic patterns in the different conditions. After discussing our results we outline how the knowledge gained may aid in the design of more natural human-machine interactions.

Index Terms: multiparty dialogue, human-computer interaction, turntaking

Figure 1: *Examples of chat (top) and chunk (bottom) phases in two stretches from the 5-party conversation analysed in this work. Each row denotes the activity of one speaker across 120 seconds. Speech is green, and laughter is yellow on a grey background (silence).The chat frame, taken at the beginning of the conversation, can be seen to involve shorter contributions from all participants with frequent laughter. The chunk frame shows longer single speaker stretches.*

1. References

[1] E. Ventola, "The structure of casual conversation in English," *Journal of Pragmatics*, vol. 3, no. 3, pp. 267–298, 1979.

[2] S. Eggins and D. Slade, *Analysing casual conversation.* Equinox Publishing Ltd., 2004.

[3] E. Gilmartin, F. Bonin, L. Cerrato, C. Vogel, and N. Campbell, "What's the game and who's got the ball? Genre in spoken interaction," 2014.

Using Visual Speech Information for Speech Enhancement

Danny Websdale, Ben Milner

University of East Anglia, United Kingdom

d.websdale@uea.ac.uk, b.milner@uea.ac.uk

This work is concerned with using neural networks for estimating ratio masks within a speech enhancement framework. We initially examine audio-only applications, where recent work has focused on improving the temporal modelling nature of the network, moving from standard feed-forward neural networks (DNN) into recurrent neural networks (RNN). Recurrent neural networks have been improved by moving from single forward directional using Long Short-term Memory (LSTM) cells into both forward and backward bidirectional LSTMs (BiLSTM).

We first explore improving this temporal recurrent network model by initially comparing the currently used BiLSTM system with the more recently proposed BiGRU alternative to BiLSTM. We then expand on this model by proposing a recurrent feed-forward hybrid (RNN-DNN) system which introduces a connection from the input of the recurrent network with its output shown in Figure 1.

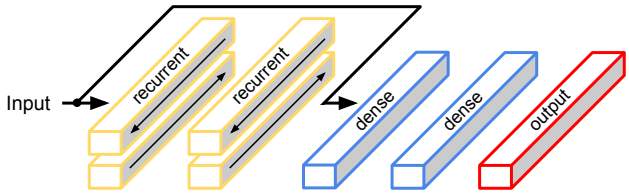


Figure 1: *RNN-DNN architecture.*

Furthermore, we examine the effect of supplementing the audio features used in mask estimation with visual speech information. Visual speech is known to be robust to noise although not necessarily as discriminative as audio features, particularly at higher signal-to-noise ratios (SNR). Comparisons are made between using a traditional visual feature extraction technique, active appearance models (AAM), with convolutional neural networks (CNN) which allow the network to learn its own features from the raw visual frame. The CNN and temporal network are trained simultaneously in a single model allowing full backpropagation throughout the model. An overview of the proposed system is shown in Figure 2.

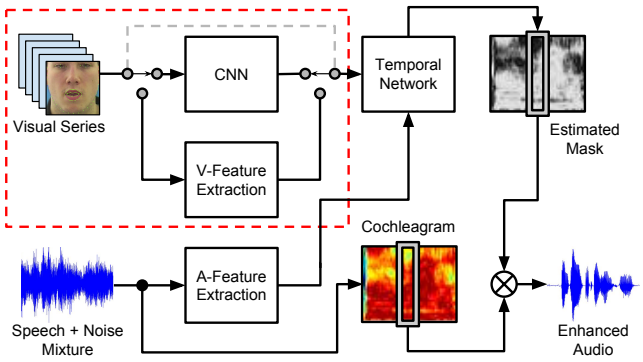


Figure 2: Overview of the speech separation system.

Evaluations of the proposed bimodal convolutional recurrent neural network architectures are carried out using a single speaker from the GRID dataset. Experiments find significant improvements combining visual with audio over audio-only systems, particularly at low SNR. Improvements are also found with the proposed RNN-DNN architecture over RNN only across both audio-only and audio-visual systems. For visual features experiments suggest using CNNs over AAMs even in this favourable single speaker configuration where AAMs perform well.

An avatar-based system for identifying individuals likely to develop dementia

Bahman Mirheidari[1], Daniel Blackburn[3], Kirsty Harkness[4], Traci Walker[5], Annalena Venneri[3,6], Markus Reuber[7], and Heidi Christensen[1,2]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK

[2]Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, Sheffield, UK

[3]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK

[4]Department of Neurology, Royal Hallamshire Hospital, Sheffield, UK

[5]Department of Human Communication Sciences, University of Sheffield, Sheffield, UK

[6]IRCCS Fondazione Ospedale San Camillo, Venice, Italy

[7]Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, UK

{bmirheidari2,d.blackburn,traci.walker,a.venneri,m.reuber, heidi.christensen}@sheffield.ac.uk, kirsty.harkness@sth.nhs.uk

Abstract

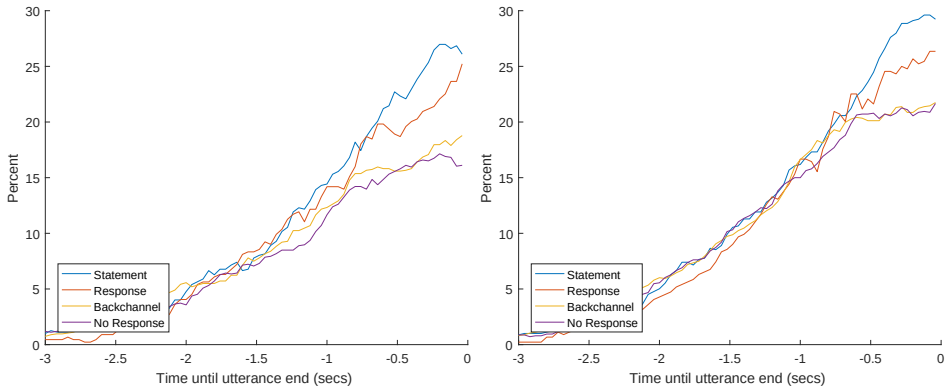
This work focuses on developing an automatic dementia screening test based on patients' ability to interact and communicate - a highly cognitively demanding process where early signs of dementia can often be detected. Such a test would help general practitioners, with no specialist knowledge, make better diagnostic decisions as current tests lack specificity and sensitivity. We investigate the feasibility of basing the test on conversations between a 'talking head' (avatar) and a patient and we present a system for analysing such conversations for signs of dementia in the patient's speech and language. Previously we proposed a semi-automatic system that transcribed conversations between patients and neurologists and extracted conversation analysis style features in order to differentiate between patients with progressive neurodegenerative dementia (ND) and functional memory disorders (FMD). Determining who talks when in the conversations was performed manually. In this study, we investigate a fully automatic system including speaker diarisation, and the use of additional acoustic and lexical features. Initial results from a pilot study are presented which shows that the avatar conversations can successfully classify ND/FMD with around 91% accuracy, which is in line with previous results for conversations that were led by a neurologist.

Towards predicting dialog acts from previous speakers' non-verbal cues

Matthew Roddy and Naomi Harte

ADAPT Centre, School of Engineering
Trinity College Dublin, Ireland

In this analysis we look at non-verbal speaker behaviors that can be used to predict the subsequent speaker's dialog act. The application of these predictions is to inform incremental approaches to the design of conversational agents (robots and avatars). It has been established in the literature that there are patterns of non-verbal behavior observed in human dyadic converations that are used to regulate turn-taking. These social signals enable humans to predict aspects of what the speaker is going to say before they finish their turn, allowing listeners to prepare their responses in advance of the turn-switch. We identify four types of these non-verbal signals: inner eyebrow movement, outer eyebrow movement, blinks, and gaze at interlocutor. We define three categories of dialog acts: response, statement, and backchannel. In addition we define a fourth category, no response, which is not a dialog act but is a relevant category for agent interactions. We perform a frame-based analysis for the presence of each of the non-verbal signals in a speaker directly preceding the different categories of dialogue acts of their speaking partner. We use the IFADV corpus, as well as automatic methods of extracting facial action units from video signals. Figure 1 shows that there are a number of patterns that can be used to predict the subsequent dialog act. For example, in both of the brow features the statement and response percentages diverge from the no-response and backchannel graphs around the -0.7 second mark. This could imply that the intention to take a speaking turn is manifested by brow movement somewhere in that region. The graphs show that non-verbal behaviors are a potentially rich source of information to inform dialog decisions.



(a) AU01 Inner Brow Raiser

(b) AU02 Outer Brow Raiser

(c) AU45 Blink

(d) Gaze at addressee

Figure 1: The percentage of frames in which each non-verbal feature is observed in speakers leading up to the end of their utterance. The different dialogue acts correspond to those of the subsequent speaker.

COMBINING INFORMATION FROM MULTIPLE SOURCES FOR ASV-ANTISPOOFING

Bhusan Chettri and Bob L. Sturm

Queen Mary University of London, United Kingdom

Automatic speaker verification (ASV) aims to verify whether a spoken utterance belongs to the claimed speaker or not. With the recent advancement in technology, ASV systems have found increasing demand and use for biometric authentication in various sectors such as security firms, banks and mobile phones. ASV systems, however, are not 100% secure and can be highly vulnerable to spoofing attacks. Attempting to fool a biometric system is called "spoofing". Mimicry, Text-to-Speech, Voice conversion and Replay attacks are four commonly used techniques for spoofing an ASV system. The "replay" attack is the simplest spoofing approach to bypass an automatic speaker verification system. This kind of attack involves replaying pre-recorded speech of an enrolled speaker. Figure 1 illustrates the difference between a genuine and a replayed speech and show the task of an spoofing countermeasure.

The first ASV spoofing challenge was held in 2015 that focused on the TTS and VC attacks. The second edition of the challenge, ASVSpoof 2017, however, focuses on text-dependent replay attack detection 'in the wild' with varying acoustic conditions [1]. Given a recorded speech utterance s, the main goal of the challenge is to determine if it is a genuine speech.

In this paper, we describe our contribution to the 2017 ASVspoof challenge. First, we study the effectiveness of 8 different features (MFCC, IMFCC, LFCC, RFCC, LPCC, SCMC, SSFC, CQCC) and machine learning approaches for the automatic detection of replayed speech on the ASVspoof 2017 challenge dataset. These features have shown good performance on ASVspoof 2015 database but that was focused on VC and TTS attacks. Replay attacks, however, are in principle very different from these artificial speech attacks: while they have the objective - to bypass an ASV system - they are acoustically different. Therefore, it is not obvious whether one should expect the extensive prior results reported in [2] on the ASVspoof 2015 data to generalize at all to detection of replay attacks. Thus, our primary scientific contribution is to thoroughly assess performance of these features on the ASVSpoof 2017 datasets. Second, we investigate how spoofing detection performance improves by combining multiple sources through score-level fusion. We compare 5 different fusion approaches: linear logistic regression (LS), score averaging, least squares (LS), LASSO and K-nearest neighbor (KNN).

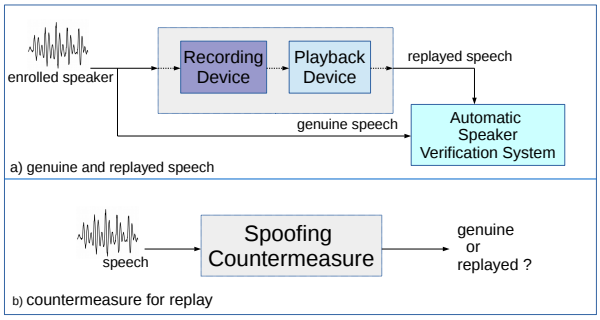


Fig. 1. Replayed, genuine speech and spoofing countermeasure illustration.

Our experimental results suggest that static features are inadequate for replay spoofing detection and derivative features (D, DA) seem to capture discriminative cues between genuine and replayed speech and show best performance. SCMC, RFCC and IMFCC are the top three performing features. Linear logistic regression gave the best fusion performance and KNN the worse among 5 approaches we studied. Our DA feature fusion system based on LR give the best performance of 10.98% when pooled data is used for model training. This, further reduces to 10.52% when only development data is used for training GMM. Although, our system does not outperform the deep learning based winning system performance (6.73%) [1], our computationally lightweight GMM system outperforms the two baseline (30.6% and 24.77%) by a large margin.

Our future work aims at exploring fusion of hand-crafted features (SCMC, IMFCC) and learned features from Deep Neural Network, so called tandem features, for replay spoofing detection.

1. REFERENCES

[1] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of audio replay attack detection in the wild," in *Interspeech*, 2017.

[2] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Interspeech 2015*, 2015.

MagPhase Vocoder: Magnitude and Phase Analysis/Synthesis for Statistical Parametric Speech Synthesis

Felipe Espic, Cassia Valentini-Botinhao, and Simon King

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

felipe.espic@ed.ac.uk, cvbotinh@inf.ed.ac.uk, Simon.King@ed.ac.uk

Vocoder code and samples:
http://www.felipeespic.com/magphase

Abstract

We propose a simple new representation for the FFT spectrum tailored to statistical parametric speech synthesis. It consists of four feature streams that describe magnitude, phase and fundamental frequency using real numbers. The proposed feature extraction method does not attempt to decompose the speech structure (e.g., into source+filter or harmonics+noise). By avoiding the simplifications inherent in decomposition, we can dramatically reduce the "phasiness" and "buzziness" typical of most vocoders. The method uses simple and computationally cheap operations and can operate at a lower frame rate than the 200 frames-per-second typical in many systems. It avoids heuristics and methods requiring approximate or iterative solutions, including phase unwrapping.

Two DNN-based acoustic models were built - from male and female speech data - using the Merlin toolkit. Subjective comparisons were made with a state-of-the-art baseline, using the STRAIGHT vocoder. In all variants tested, and for both male and female voices, the proposed method substantially outperformed the baseline. We provide source code to enable our complete system to be replicated.

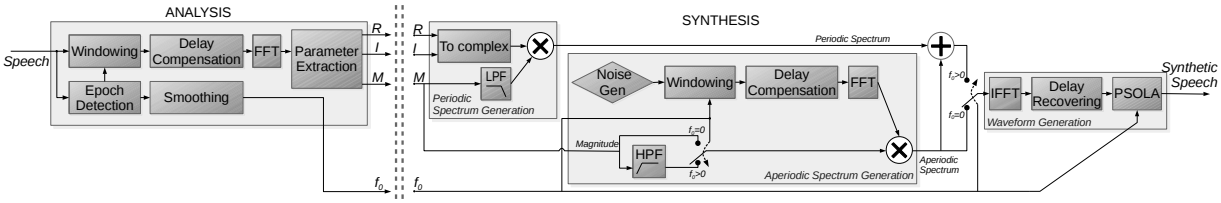


Figure 1. Diagram of the analysis and synthesis processes of the MagPhase vocoder. Four features: f_0 , *M*, *R*, and *I* are extracted to synthesise speech.

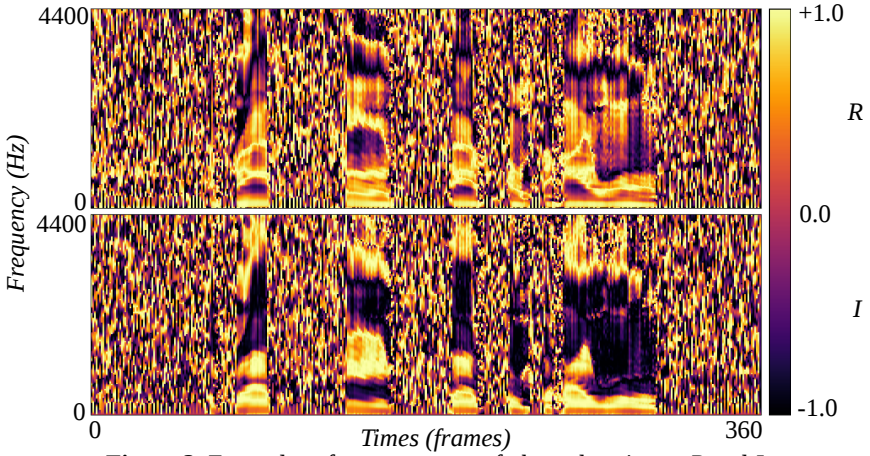


Figure 2. Examples of spectrograms of phase descriptors *R* and *I*.

Phrase break prediction for long-form reading TTS: exploiting text structure information

Viacheslav Klimkov, Adam Nadolski, Alexis Moinet, Bartosz Putrycz, Roberto Barra-Chicote, Thomas Merritt, Thomas Drugman

Amazon.com

vklimkov,anadolsk,amoinet,bartosz,rchicote,thommer,drugman@amazon.com

Abstract

Phrasing structure is one of the most important factors in increasing the naturalness of text-to-speech (TTS) systems, in particular for long-form reading. Most existing TTS systems are optimized for isolated short sentences, and completely discard the larger context or structure of the text.

This paper presents how we have built phrasing models based on data extracted from audiobooks. We investigate how various types of textual features can improve phrase break prediction: part-of-speech (POS), guess POS (GPOS), dependency tree features and word embeddings. These features are fed into a bidirectional LSTM or a CART baseline. The resulting systems are compared using both objective and subjective evaluations. Using BiLSTM and word embeddings proves to be beneficial.

Index Terms: speech synthesis, TTS, BiLSTM, long form reading, phrasing, respiratory pauses, audiobooks

Speech intelligibility in cars: the effect of speaking style, noise and listener age

Cassia Valentini Botinhao[1], *Junichi Yamagishi*[1,2,3]

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

cvbotinh@inf.ed.ac.uk, jyamagis@inf.ed.ac.uk

1. Abstract

The use of speech interfaces in cars over visual displays increases safety as drivers are not required to look away from the road [1]. This is particularly relevant for older adults as this group of drivers tend to focus more on the road, spending less time looking at in-vehicle displays [2]. Thus, using speech as a way to provide information to older drivers seems like a good choice. However, older drivers are more likely to experience age-related and noise-induced elevations of auditory thresholds, as well as increased mental workload.

Elderly listeners experience difficulties processing speech, particularly in noise and under stress. This holds true even when there is no evidence of abnormal hearing thresholds [3]. It has been argued [4] that this could be due to temporal resolution loss caused by disrupted neural connections that happen with age. According to [4], this kind of loss appears as a decline in the processing of the slowly varying envelope (as observed by measuring performance on gap detection tasks [5]) and the processing of the fine structure envelope (as observed by measuring performance on frequency modulation detection [6], pitch discrimination [7], inter-aural phase and time difference detection tasks [8]). This impacts speech understanding in noise, but particularly in fluctuating noises, as good temporal acuity is necessary for a listener to take advantage of the relatively silent gaps in the noise. In [5] it was found that word recognition in competing babble correlated significantly with temporal resolution and age, but not with absolute hearing loss. In [9], however, it was found that in stationary noise, hearing loss significantly contributed to explaining differences in speech reception and no other predictors (age, temporal acuity) seemed to contribute.

It is possible to modify speech in such a way that the mixture of speech and noise is more intelligible for the listener without an overall level increase. One could for instance modify speech produced in quiet conditions by promoting acoustic changes observed in speaking styles that are more intelligible in noise conditions, such as clear speech (produced with the intent to counter adverse listening conditions) and Lombard speech (produced in noise). A study [10] that investigated the driving performance of a group of university students found that navigation systems with dominant voices (faster speech rate, higher amplitude and pitch, more pitch variation and dominant messages) lead drivers to follow instructions better. Both clear and Lombard speech have been shown to increase intelligibility for older adults [11], although some studies did not find the clear speech benefit for hearing impaired older adults [12].

In this work [13], we are interested in finding which driving conditions and which speaking styles are more intelligible for older adults. For this purpose we have recorded a database of a variety of voices and speaking styles in many different driving conditions. We then performed a listening experiment on a selected portion of the data to gather intelligibility scores from young (below 30 years of age) and older (above 50) adults. We found that older participants intelligibility scores were lower for competing speaker and background music. Results also indicate that clear and Lombard speech were significantly more intelligible than plain speech for both age groups.

Acknowledgements: The work presented in this paper was partially supported by the TOYOTA motor cooperation.

2. References

[1] A. Barón and P. Green, "Safety and usability of speech interfaces for in-vehicle tasks while driving: A brief literature review," University of Michigan, Transportation Research Institute, Tech. Rep., 2006.

[2] I.-M. Jonsson, M. Zajicek, H. Harris, and C. Nass, "Thank you, I did not see that: in-car speech based information systems for older adults," in *Proc. CHI*. ACM, 2005, pp. 1953–1956.

[3] K. S. Helfer and R. L. Freyman, "Aging and speech-on-speech masking," *Ear and Hearing*, vol. 29, no. 1, p. 87, 2008.

[4] T. Schoof and S. Rosen, "The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners," *Frontiers in Aging Neurosc.*, vol. 6, p. 307, 2014.

[5] K. B. Snell, "Age-related changes in temporal gap detection," *J. Acoust. Soc. Am.*, vol. 101, no. 4, pp. 2214–2220, 1997.

[6] N.-J. He, J. H. Mills, and J. R. Dubno, "Frequency modulation detection: effects of age, psychophysical method, and modulation waveform," *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. 467–477, 2007.

[7] C. Füllgrabe, "Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss," *Am. J. of Audiology*, vol. 22, no. 2, pp. 313–315, 2013.

[8] J. H. Grose and S. K. Mamo, "Processing of temporal fine structure as a function of age," *Ear and Hearing*, vol. 31, no. 6, p. 755, 2010.

[9] E. L. George, A. A. Zekveld, S. E. Kramer, S. T. Goverts, J. M. Festen, and T. Houtgast, "Auditory and nonauditory factors affecting speech reception in noise by older listeners," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2362–2375, 2007.

[10] M. Jonsson and N. Dahlbäck, "In-car information systems: Matching and mismatching personality of driver with personality of car voice," in *Proc. ICHCI*, 2013, pp. 586–595.

[11] M. Fitzpatrick, J. Kim, and C. Davis, "Auditory and auditory-visual lombard speech perception by younger and older adults." in *Proc. AVSP*, 2013, pp. 105–110.

[12] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 259–271, 2002.

[13] C. Valentini-Botinhao and J. Yamagishi, "Speech intelligibility in cars: the effect of speaking style, noise and listener age." in *Proc. Interspeech*, 2017.

Towards Lipreading Sentences with Active Appearance Models

George Sterpu, Naomi Harte

Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

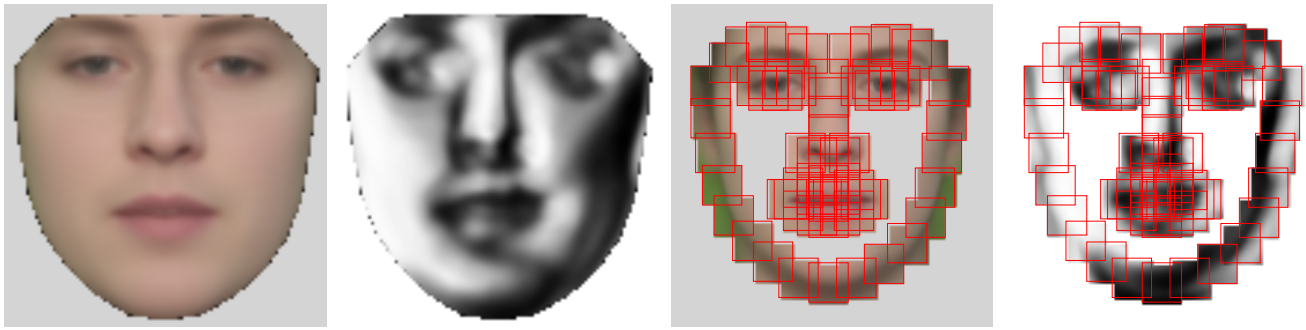
sterpug@tcd.ie, nharte@tcd.ie

1. Abstract

Understanding the visual modality of speech, widely known as lipreading, yields the benefit of complementing and supplementing the acoustic one, otherwise dominant in clean conditions. Most attempts at lipreading have been performed on small vocabulary tasks, due to a shortfall of appropriate audio-visual datasets.

In this work we use the publicly available TCD-TIMIT database, designed for large vocabulary continuous audio-visual speech recognition. We compare the viseme recognition performance of the most widely used features for lipreading, Discrete Cosine Transform (DCT) and Active Appearance Models (AAM), in a traditional Hidden Markov Model (HMM) framework. We also exploit recent advances in AAM fitting. We found the DCT to outperform AAM by more than 6% for a viseme recognition task with 56 speakers. The overall accuracy of the DCT is quite low (32-34%). We conclude that a fundamental rethink of the modelling of visual features may be needed for this task.

Active Appearance Models (AAM), introduced in [1] and streamlined in [2], are state-of-the-art techniques for deformable object modeling. The robustness of AAMs has greatly improved since these early publications via several factors: better fitting algorithms [3], feature-based image descriptors [4] and patch models [5] (portrayed in Figurge 1). These improvements are fairly recent, yet remarkable efforts have been invested to make them available in an open-source project [6].



(a) *Holistic no-op* (b) *Holistic SIFT* (c) *Patch no-op* (d) *Patch SIFT*

Figure 1: *Overview of AAM types by warp and feature used.*
The Patch *models are evaluating local neighborhoods of the landmarks instead of the entire appearance. The SIFT descriptors are robust alternatives to raw pixel intensities, where no additional operation is applied (no-op).*

A first finding is that AAM features do not outperform the DCT ones in an identical recognition framework. This is the most comprehensive comparison between DCT and state-of-the-art AAM that we are aware of.

An important conclusion about AAMs is that the *Patch* models, especially when combined with SIFT image descriptors, are able to achieve a much higher fitting and implicitly recognition accuracy than the traditional *Holistic* ones that have been used so far in lipreading. As shown in [5], their robustness is conspicuous when trained on unconstrained in-the-wild faces, making them more suitable candidates for realistic lipreading scenarios.

Index Terms: Visual Speech Recognition, DCT, AAM, Large Vocabulary, TCD-TIMIT

2. References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," 1998, pp. 484–498.

[2] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[3] J. Alabort-i Medina and S. Zafeiriou, "A unified framework for compositional fitting of active appearance models," *International Journal of Computer Vision*, vol. 121, no. 1, pp. 26–64, 2017.

[4] E. Antonakos, J. A. i Medina, G. Tzimiropoulos, and S. P. Zafeiriou, "Feature-based lucas-kanade and active appearance models," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2617–2632, Sept 2015.

[5] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1851–1858.

[6] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 679–682.

Notes

Notes