PROGRESS ON PHONEME RECOGNITION WITH A CONTINUOUS-STATE HMM

P. Weber, L. Bai, S. M. Houghton, P. Jančovič and M. J. Russell

School of EESE, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
dr.philip.weber@ieee.org,{s.houghton,lxb190,p.jancovic,m.j.russell}@bham.ac.uk

ABSTRACT

Recent advances in automatic speech recognition have used large corpora and powerful computational resources to train complex statistical models from high-dimensional features, to attempt to capture all the variability found in natural speech. Such models are difficult to interpret and may be fragile, and contradict or ignore knowledge of human speech production and perception. We report progress towards phoneme recognition using a model of speech which employs very few parameters and which is more faithful to the dynamics and model of human speech production. Using features generated from a neural network bottleneck layer, we obtain recognition accuracy on TIMIT which compares favourably with traditional models of similar power. We discuss the implications of these results for recognition using natural features such as vocal tract resonances and spectral energies.

Index Terms— Continuous-State HMM, phoneme recognition, neural network, bottleneck features, formants.

1. INTRODUCTION

Recent significant progress in automatic speech recognition has been achieved predominantly using statistical methods such as Deep Neural Networks (DNNs [2]) to model distributions over speech features. Very large corpora and powerful computational resources (e.g. [3]) enable training of models with many parameters, from rich high-dimensional features.

This approach assumes training and test data drawn from the same distribution, and aims to model all the expected variability in speech from the target domain, to reduce the risk of encountering novel patterns in production. With enough training data, although the model is over-trained, the empirical distributions of training and test samples will be close enough for lack of generalisation not to be a problem.

The cost can be inflexibility when applied to speech from outside the target domain. This is demonstrated by the fact that research is active into recognition of accented (e.g. [4]), children's [5] or dysarthric [6] speech, as well as training for low resource languages (e.g. [7, 8]), speech in noise [9], and model adaptation [10]. Adaptation is hampered by difficulty in interpreting large statistical models, the structures learned, and roles and behaviours of elements of the models.

Data-driven training also often ignores or contradicts findings of research on human speech production and perception. Speech is generated by the relatively slow, constrained and smooth movement of a small number of articulators in the vocal tract. Features are therefore strongly correlated in time and typically exhibit smooth, slowly-varying dynamics. It has long been argued [11, 12] that speech features thus lie on a low-dimensional data manifold embedded in high-dimensional acoustic space. DNN work tacitly acknowledges this through dimensionality reducing transforms applied to the input features (e.g. [13]), and non-linear reductions such as relatively low-dimension bottleneck layers [14].

Jansen [12] argues that modelling this manifold directly would allow recognition to be carried out closer to the original intent, perhaps therefore more robustly to noise and variability. It would also allow the dynamics of the signal to be taken into account. Segmental [15, 16] and dynamical [17, 18] models attempt to model the dynamics of speech more faithfully, but are hampered by computational complexity.

The Continuous State HMM (CS-HMM) [19] can be cast as a type of segmental model [20]. Its iterative computations avoid some of these problems, and it can be trained on limited data of low dimensionality. Variants have been applied to voiced sounds [21] with formant-type features, and unvoiced sounds [23] using spectral energy features.

We plan to integrate these models into a full recogniser which would probabilistically combine hypotheses from multiple models and heterogeneous views on the data (see e.g. [24]). Questions remain, including how to automatically choose appropriate features for each observation, and combine scores from different feature spaces. As an intermediate step, in this paper we side-step these questions by building on work reported in [25] to automatically derive a low-dimensional representation of speech, valid for all speech sounds (as hypothesised by Jansen et al. [12]), and faithful to the assumptions of the CS-HMM. We report promising phoneme recognition results using these bottleneck features.

2. CONTINUOUS-STATE HMM

The CS-HMM model of speech [19, 21] aims to reflect speech structure and dynamics more faithfully than conventional HMMs, reducing the assumptions that speech is a piece-wise

stationary process with temporally independent observations, and improving duration modelling. The model is inspired by the Holmes, Mattingly and Shearme (HMS) [29] dwell-transition model of speech, in which stationary dwells represent phoneme targets and transitions the smooth movement between them, corresponding to the smooth movement of the human articulators. Given N phonemes, we estimate an inventory of phoneme 'canonical' targets. Let $\boldsymbol{\theta}_\varphi$ be the target feature vector for phoneme φ. Realisation $\boldsymbol{r}_\varphi$ of φ will vary, e.g. with speaker and context. We assume this variation to be Gaussian around the target, with covariance A. We assume also observation $\boldsymbol{y}_t$ at time t to be drawn from a Gaussian around $\boldsymbol{r}_\varphi$ with covariance E. In this work, these are global covariances, but they could be estimated per-phoneme. Thus

$$\boldsymbol{r}_\varphi \sim \mathcal{N}(\boldsymbol{\theta}_\varphi, \mathrm{A}), \qquad \boldsymbol{y}_t \sim \mathcal{N}(\boldsymbol{r}_\varphi, \mathrm{E}). \tag{1}$$

The trained system contains a model $\boldsymbol{\theta}_\varphi$ per phoneme, two covariances A and E, and a timing model, which in this work simply allows uniformly distributed dwell and transition durations over a specified range. These at most several hundred parameters (see Tables 1 and 2), are estimated from data as described in the next section, as is a language model.

Recognition uses a sequential branching algorithm to recover the most likely sequence of alternating dwells and transitions, the times of changes between them, and the sequence of phonemes which generated them. Hypotheses are maintained for all possible trajectories, pruning the least likely for computational efficiency. Each hypothesis maintains a 'state' consisting of continuous components $\boldsymbol{x}_t$ and discrete components $\boldsymbol{d}_t$, maintained as a Baum-Welch alpha value,

$$\alpha_t(\boldsymbol{x}, \boldsymbol{d}) = K_t n(\boldsymbol{x} - \boldsymbol{\mu}_t, \mathrm{P}_t), \tag{2}$$

which stores information about an infinite set of explanations of the data, as a scaled Gaussian. It represents the hypothesis' belief of the current realisation, given the observations seen, the current hypothesised phoneme and phonetic history, and duration of the current dwell or transition.

On each observation, hypotheses are split to account for the possibilities of continuing in the current dwell or transition, or changing from dwell to transition or vice versa. A distinguishing feature of the dwell-transition CS-HMM is that continuity is preserved across the segment boundaries.

2.1. Training Procedure

To estimate parameters we use a Viterbi alignment procedure [21]. Initial estimates are obtained using the TIMIT [1] transcribed phoneme boundaries, to identify which features belong to which phoneme. This assumes that dwells extend between these boundaries and there are no transitions. We use all 'non-SA' utterances for training. Each utterance is then decoded with the CS-HMM decoding algorithm using a strict language model, the true sequence of phonemes for the utterance. The most likely hypothesis returned will include a first estimate of boundaries between dwells and transitions. From these improved boundaries, an improved set of phoneme targets $\boldsymbol{\theta}_\varphi$ and realisation covariance A can be estimated from the features now marked as dwell phases. Observation covariance E is estimated from both dwells and transitions. Decoding is then repeated with the new inventory, until convergence (boundaries and parameter estimates no longer change).

3. FEATURES

In this section we briefly outline the derivation of bottleneck features and describe other features used in the experiments.

3.1. Low-Dimensional Bottleneck Features (BNs)

We obtain bottleneck features using a neural network classifier as described by Bai et al. [25]. Log Mel frequency filterbanks (26 channels) were obtained from TIMIT audio sampled at 16kHz, analysed using a 25ms Hamming window with 10ms frame rate, normalised to zero mean and unit variance over the training set. Windows of 11 features (central $\pm$ 5 frames) were input to a 5-layer multi-layer perceptron giving a 286 neuron input layer. Hidden layers contained sigmoid-activation neurons, 512 in layers 2 and 4, with a 3 or 9 neuron bottleneck in layer 3. Using the Theano toolkit [27], the network was trained discriminatively using Stochastic Gradient Descent with the cross-entropy error criterion, to predict phoneme posterior probabilities from the 'standard' 49 set of TIMIT phonemes [26]. Training was halted at the soonest of increasing validation set error, or at 3000 epochs. We used 90% of TIMIT 'Train' for training, 10% for validation.

We generated bottleneck features for the whole of TIMIT by feeding the same input features to the trained network, and recording the activations at the bottleneck layer. Several sets of bottleneck features were obtained from networks with the above structure trained from different random initialisations.

3.2. Formants and Vocal Tract Resonances (VTRs)

The HMS model was originally described in terms of formants, the resonances of the human vocal tract as mainly observed during sonorant speech. We use Wavesurfer [30] to obtain trajectories for F_1, F_2 and F_3 from TIMIT. Formants are notoriously hard to estimate accurately, and not meaningful for all speech sounds [25, Fig. 4b)], while the underlying Vocal Tract Resonances (VTRs) manifesting as formants during sonorant speech are postulated as valid for all speech. The VTR database [22] provides VTRs for a subset of TIMIT.

3.3. Perceptually-Motivated Spectral Features

Perceptual experiments have shown that humans discriminate between unvoiced sounds largely on the basis of broadband energy between specific frequencies and of specific duration. Between such sounds, acoustic change is abrupt, so the HMS

Features	Phonemes	Dim.	Train	Test	Model	Corr	Sub	Del	Ins	Err	Acc	#Parm
39 MFCC + δ + $\delta\delta$ [25]	all	39	Train	Core Test	DS-HMM	76.2	–	–	–	29.1	70.9	1.4e7
9D Bottleneck [25]	all	9	Train	Core Test	DS-HMM	74.4	17.8	8.8	2.9	29.4	70.6	2.3e5
3D Bottleneck [25]	all	3	Train	Core Test	DS-HMM	65.0	24.2	10.8	4.1	39.1	60.9	7.6e4
3 Formant [25]	all	3	Train	Core Test	DS-HMM	49.3	32.0	18.7	8.6	59.3	40.7	7.6e4
3 Formant + δ + $\delta\delta$ [25]	all	9	Train	Core Test	DS-HMM	56.3	24.3	19.3	5.2	48.9	51.1	2.3e5
3 VTR [21]	voiced (v)	3	1 Speaker	1 Speaker	CS-HMM	39.6	31.1	29.3	2.4	62.8	37.2	85
9 Spectral Energies [23]	unvoiced (uv)	9	Train	Core Test	CS-HMM	73.1	19.5	8.2	3.2	30.8	69.2	245

Table 1. Phone % error (etc.) from previous phoneme recognition experiments. Top: 'standard' discrete tied-state triphone HMM-GMM (DS-HMM) (approx. $11,000$ models); 13 MFCCs plus deltas and delta-deltas. Centre: monophone DS-HMM comparing formants and bottleneck features. Bottom: CS-HMM, training and testing on (v) voiced phoneme sequences from a single-speaker, (uv) unvoiced phoneme sequences. All results use a bigram language model. Parameter count #Parm is for the model only, excludes LM and feature extraction.

model is not a good fit. Instead, vectors of spectral energies between perceptually-motivated frequencies can be used with a 'dwell-only' model to decode unvoiced consonants [23].

4. RESULTS

In this section we briefly review previous phoneme recognition results using bottleneck features with discrete-state HMMs (DS-HMMs), and limited experiments using the CS-HMM with 'natural' features (VTRs and spectral features).

4.1. Previous Results

Using 9-dimensional bottleneck (9D BN) features, and a 'standard' discrete-state HMM system implemented in HTK [28], phone accuracy was achieved almost equivalent to that obtained with MFCCs (Table 1, lines 1 and 2) [25]. Accuracy with the BNs was considerably better than with equivalent-dimension formant features (lines 3-5). Visualisations suggested that the BNs preserved the time dynamics of speech well, better and more consistently than formants, and should therefore be suitable for recognition with the CS-HMM. Very little improvement was seen with higher-dimension BNs.

Using a CS-HMM with VTRs, Houghton et al. [21] trained and tested on sequences of voiced sounds only, for a single TIMIT speaker. The results showed the ability of the training algorithm to learn from small amounts of training data, but the best phone accuracy was only 37.2%, using 3 VTRs and a bigram language model (Table 1, line 6). Extending to multiple speakers (Table 2, lines 4-6), errors increase significantly, suggesting that the model at present cannot account well for the variability in the VTR trajectories. The problem is even worse for formants (Table 2, lines 1-3).

Using perceptually-motivated spectral energies, Weber et al. [23] obtained 30.8% phone error on sequences of unvoiced TIMIT phonemes, trained and tested on the full TIMIT Train and Core Test. This is considerably better than obtained for voiced sounds with VTRs and suggests that these features are much less sensitive to variability between speakers.

4.2. Bottleneck Results

In the lower half of Table 2 we report phoneme recognition results for full TIMIT utterances (labelled 'all'), voiced phonemes (32 vowels, liquids, aspirates, nasals and voiced fricatives and affricates, labelled 'v') and unvoiced phonemes (17 stops, closures and unvoiced fricatives and affricates, labelled 'uv'). Models were built for the appropriate subset from the 'standard' mapping to 49 phonemes, and scored using the mapping to 40 [26]. The results reported are means from repeated experiments using BNs from neural networks trained from different random initialisations. The top part of the table gives results for formants and VTRs for comparison.

Accuracy using BNs with the CS-HMM was not quite as good as with the DS-HMM (Table 1) but the CS-HMM used several orders of magnitude fewer parameters. The BNs perform significantly better than formants and VTRs (in either model), suggesting that they successfully eliminate much of the variability in these features which the CS-HMM was unable to handle (Section 4.1 and [21]).

Curiously, using BNs %Err was higher for voiced sounds than for full utterances. It is possible that features generated by the neural network are less consistent for voiced sounds than for unvoiced. We hypothesise that this is an effect of the network training procedure (to predict phoneme posteriors) implicitly assuming a 'dwell-only' model (features stationary throughout a phoneme) rather than dwell-transition. This could also explain the lower error for unvoiced phonemes, whose features are more stationary [23] (although this may simply be due to fewer classes of unvoiced phonemes).

Finally, for unvoiced sounds, BNs performed better than spectral features, perhaps because no forced alignment was used in the previous experiments using spectral features.

5. DISCUSSION AND FUTURE WORK

The BN results are encouraging in showing for the first time that speech recognition using a CS-HMM 'segmental' model is possible given appropriate features. With 9D features, 38.1% error is not too dissimilar from the baseline MFCC result, while using significantly fewer parameters. The CS-

Features	#Phn	Corr	Sub	Del	Ins	Err (S/E)	#Parm
3 Formant	all	31.1	35.6	33.4	4.8	73.7	163
3 Formant	v	20.4	31.2	48.4	1.6	81.2	112
3 Formant	uv	31.9	33.2	34.9	4.2	72.3	67
3 VTR	all	29.2	36.2	34.6	3.7	74.6	163
3 VTR	v	29.2	37.0	33.8	3.3	74.2	112
3 VTR	uv	32.2	33.4	34.4	2.5	70.3	67
3D BN	all	55.7	30.1	14.2	3.6	47.9 (0.07)	163
3D BN	v	52.5	29.3	18.2	3.4	50.9 (0.09)	112
3D BN	uv	71.9	17.4	10.7	2.3	30.4 (0.01)	67
9D BN	all	66.9	22.9	10.2	5.0	38.1 (0.11)	535
9D BN	v	60.9	24.6	14.5	3.9	43.0 (0.01)	382
9D BN	uv	82.8	10.9	6.3	4.2	21.3 (0.25)	247

Table 2. CS-HMM phone recognition results, with formants [30] and VTRs [22] (top section) and bottleneck features (lower sections). The bottleneck results given are means over features from 5 network random initialisations, giving the standard error of the mean (S/E).

HMM used just 535 trainable parameters (per-phoneme mean features, symmetric global realisation and observation covariance matrices, and four parameters for a timing model), plus a bigram language model (2601 parameters). Minimum error was reached after 2 to 5 iterations of forced alignment, after which it began to increase, suggesting that the training algorithm is not yet optimal. With such improvements, and perhaps an improved timing model and some per-phoneme parameters, we expect to somewhat reduce the error rate.

These results are however somehow disappointing because the same problem of lack of interpretability affects our neural network-derived features, as affects recognition results from DNNs. Our aim is speech recognition using models and features interpretable in terms of human production and perception. The CS-HMM fulfils this in part; the BNs at present do not – but they outperformed 'natural' features in every case. In addition, we ignored parameters involved in generating features, approximately $180,000$ in the case of the 9D BNs (although these are only required for training the feature generator, not to train or test the recogniser).

Why do the BNs perform well? Figure 1 shows that the CS-HMM tends to fit the data well, but for the VTR and formants the inventory frequencies (green lines) are often very similar and bear little relation to the features. The phoneme inventory learned for BNs is more discriminatory and a better fit to the data, suggesting some of the variation not needed for discriminating between speech sounds has been excluded from the BNs. The formant features are also seen, as expected, to be noisy during unvoiced sounds, but the CS-HMM has tried erroneously to fit short phonemes in these regions.

Future work is planned in several directions. Firstly, to improve the ability of the CS-HMM to account for the variability in 'natural' features, for example using Vocal Tract Normalisation techniques (which can in theory be accommodated simply within the model). Secondly, to understand and improve the BNs. Previous work [21] has shown criticality of

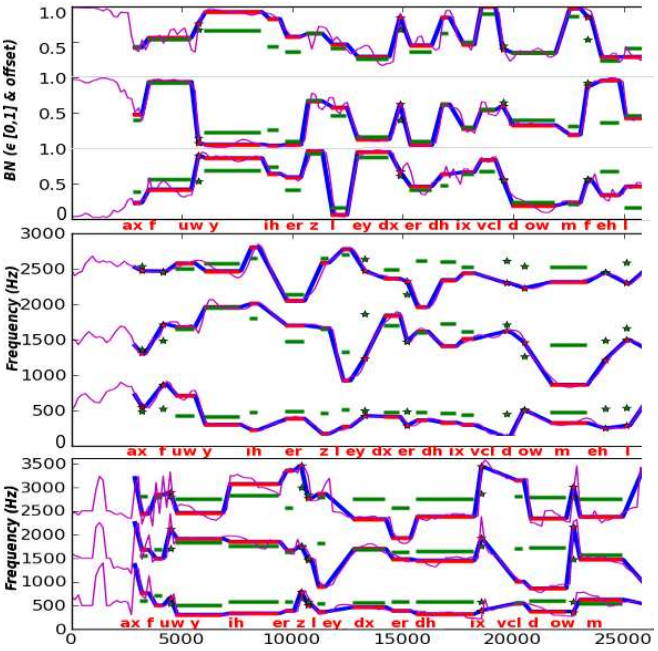

Fig. 1. Example CS-HMM recoveries (thick blue lines), showing realised dwells (red), inventory feature means (green). From top: 3D BNs (magenta) $\in [0, 1]$, offset to visualise), VTRs, formants.

an accurate inventory. The difference in performance between BNs for voiced and unvoiced sounds suggests an important link between the approach to training the networks generating the features, and the inventory which can be trained from them. Thirdly, to reduce or account for parameters involved in generating features, for example by showing that the BN generator once trained, can be applied in multiple settings.

Since the CS-HMM requires many fewer parameters, we ought to be able to train using less data than required for the DS-HMM. Alternatively, increasing the parameter count in an informed way (e.g. to encode known variants of phonemes, such as 'dark' and 'light' /l/s) may allow improvement in recognition accuracy. One advantage of the CS-HMM is that it provides a natural framework in which to incorporate such perceptual knowledge [23]. We plan therefore to investigate how such knowledge is represented in the BNs, and the effects on recognition error rates of incorporating such knowledge.

6. CONCLUSION

We reported, for the first time, TIMIT phoneme recognition results using a CS-HMM – a model of speech more faithful to human speech production – using low dimensional 'bottleneck' features, which apparently somehow capture the true dynamics of speech. We avoided the question of whether these features can be interpreted in terms of human speech production and perception. Future work will therefore focus on understanding the derived representations, and on recognition with the CS-HMM using perceptually-motivated features such as vocal tract resonances and spectral energies.

7. REFERENCES

[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa TIMIT acoustic-phonetic continuous speech corpus CD-Rom," NIST, Tech. Rep., 1990.

[2] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Trans. Audio, Speech, and Language Process.*, 20(1), pp. 14–22, 2012.

[3] O. Kapralova, J. Alex, E. Weinstein, P. J. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," in *Proc. Interspeech, Singapore*, 2014, H. Li, H. M. Meng, B. Ma, E. Chng, and L. Xie, Eds., pp. 2083–2087.

[4] M. Najafian, A. DeMarco, S. J. Cox, and M. J. Russell, "Unsupervised model selection for recognition of regional accented speech," in *Proc. Interspeech, Singapore*, 2014, H. Li, H. M. Meng, B. Ma, E. Chng, and L. Xie, Eds., pp. 2967–2971.

[5] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," in *Proc. IEEE SLT, South Lake Tahoe, NV, USA*, 2014, pp. 135–140.

[6] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, 27(6), pp. 1147–1162, 2013.

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. ICASSP 2012*, Kyoto, Japan, pp. 4269–4272.

[8] F. Grézl, M. Karafiát, and K. Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. ICASSP 2014*, Florence, Italy, pp. 7654–7658.

[9] S. H. Mallidi, T. Ogawa, K. Vesely, P. S. Nidadavolu, and H. Hermansky, "Autoencoder based multi-stream combination for noise robust speech recognition," in *Proc. Interspeech 2015*, Dresden, Germany, pp. 3551-3555.

[10] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA*, pp. 366–369.

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

[12] A. Jansen and P. Niyogi, "Intrinsic spectral analysis," *IEEE Trans. Signal Process.*, 61(7), pp. 1698–1710, 2013.

[13] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech 2013*, Lyon, France, pp. 2345–2349.

[14] L. Lu and S. Renals, "Probabilistic linear discriminant analysis with bottleneck features for speech recognition," in *Proc. Interspeech 2014*, Singapore, H. Li, H. M. Meng, B. Ma, E. Chng, and L. Xie, Eds., pp. 910–914.

[15] J. Bridle, L. Deng, J. Picone, H. Richards, J. Ma, T. Kamm, M. Schuster, S. Pike, and R. Regan, "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," CSLP JHU Tech. Rep., 1998.

[16] M. Russell and W. Holmes, "Linear trajectory segmental HMMs," IEEE Signal Process. Lett., 4(3), pp. 72–4, 1997.

[17] H. Richards and J. Bridle, "The HDM: a segmental hidden dynamic model of coarticulation," in Proc. ICASSP 1999, Piscataway, NJ, USA, vol 1, pp. 357–60.

[18] J. Frankel and S. King, "Speech recognition using linear dynamic models," *IEEE Trans. Audio, Speech, Language Process.*, 15(1), pp. 246–256, 2007.

[19] C. J. Champion and S. M. Houghton, "Application of Continuous State Hidden Markov Models to a classical problem in speech recognition," *Computer Speech and Language*, 36(1), pp347–364, 2016.

[20] S. M. Houghton and C. J. Champion, "Inductive implementation of segmental HMMs as CS-HMMs," in *Proc. Interspeech 2015*, Dresden, Germany, pp. 776-780.

[21] S. M. Houghton, C. J. Champion, and P. Weber, "Recognition of voiced sounds with a continuous state HMM," in *Proc. Interspeech 2015*, Dresden, Germany, pp. 523–527.

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

[23] P. Weber, C. Champion, S. Houghton, P. Jančovič, and M. Russell, "Consonant Recognition with Continuous-State Hidden Markov Models and Perceptually-Motivated Features," *Proc. Interspeech 2015*, Dresden, Germany, pp. 1893-1897.

[24] H. Hermansky, "Multistream recognition of speech: Dealing with unknown unknowns," *Proc. IEEE*, 101(5), pp. 1076–1088, 2013.

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

[26] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using Hidden Markov models," *IEEE Trans. Acoustics, Speech, and Signal Process.*, 37(11), pp. 1641–1648, 1989.

[27] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. the Python for Scientific Computing Conference (SciPy)*, 2010.

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

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

[30] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proc. Interspeech 2000*, Beijing, pp. 464–467.