TRAJECTORY ANALYSIS OF SPEECH USING CONTINUOUS STATE HIDDEN MARKOV MODELS

P. Weber, S. M. Houghton, C. J. Champion, M. J. Russell and P. Jančovič

School of EECE, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

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

ABSTRACT

Many current speech models used in recognition involve thousands of parameters, whereas the mechanisms of speech production are conceptually very simple. We present and evaluate a new continuous state probabilistic model (CS-HMM) for recovering dwell-transition and phoneme sequences from dynamic speech production features. We show that with very few parameters, these features can be tracked, and phoneme sequences recovered, with promising accuracy.

Index Terms— Speech Analysis, Probabilistic Model, Dynamic Features, Continuous State Hidden Markov Model

1. INTRODUCTION

The mechanisms of speech production are conceptually very simple. A small number of speech articulators move together to generate all the sounds found in speech. This premise has been used to build rule-based speech synthesis systems using few parameters, which when carefully tuned produce synthesized speech approaching natural speech in quality [1, 2]. The motion of these articulators is continuous, therefore a large part of speech consists in smooth motion of the acoustic features from properties defined by one sound to another. Acoustic feature vectors representing speech are therefore not time-independent, a fact indicated by the smooth formant tracks seen in spectrograms [2, 3, 4].

In contrast, conventional speech recognition models assume that speech can be represented by a sequence of discrete, independent events, such as the production of phonemes. In a typical system (e.g. [5]), Hidden Markov Models with Gaussian Mixture output distributions (GMM-HMMs) model each phoneme in its context of neighbouring phonemes. Speech is coded as Cepstral feature vectors (MFCCs [6]), plus a spectral energy coefficient, plus time derivatives. This results in many thousands of parameters, which attempt to implicitly model the variability and dynamics of speech [3], although parameters of similar states may be tied in an effort to reduce the total parameters in the model. Coded speech vectors are assumed to be in general time-independent, with only some approximation to local dependencies through the use of derivative features.

Recent advances in speech recognition have used Deep Belief Networks (DBNs). These multi-layer neural networks have increased the number of parameters still further and rely on ever larger corpora to estimate these parameters. While DBNs have improved speech recognition accuracy, no attempt has been made to reconcile the fundamental differences between speech production and recognition models.

Attempts have been made to define more faithful models of speech. Hidden Semi-Markov models [7], segmental HMMs [8], trajectory models, e.g. [3, 9, 10], intermediate state models [11] and Gaussian process dynamical models [12] variously relax assumptions of temporal independence or static behaviour, e.g. outputting sequences of feature vectors from each state, or modelling speech dynamics directly in an articulatory or formant-based representation. Dynamic systems approaches [3, 4, 10, 13, 14] aim to account for variability and coarticulation by modelling the evolution of features more closely related to speech production, mapped (typically non-linearly) to the acoustics. Other work tries to directly recover real articulatory features for speech recognition, possibly in conjunction with the acoustic data (e.g. [15]). Problems with these models have included computational inefficiency, and difficulty extracting adequate quality features.

We describe and evaluate a new parsimonious model for speech analysis, inspired by the Holmes-Mattingley-Shearme model [16] and in the spirit of early work due to Bridle, Paliwal and others [17, 18] which proposed modelling articulation as a series of target frequencies (dwell phases) connected by transitions. Our continuous-state hidden Markov model (CS-HMM) is a variant of the Hidden Gaussian Markov Models described by Ainsleigh *et al.* [19], but to the best of our knowledge this is the first time that such a model has been applied to speech. It is computationally efficient, probabilistic over the features and feature spaces with which it works, and minimises unwarranted independence assumptions.

After introducing the theory in Section 2, in Section 3 we show results of experiments applying the model to several utterances from the TIMIT corpus [20]. From these early results we find that even using very simple features and few parameters, this new model can faithfully track the underlying dynamics of speech, and holds promise as the foundation for future work to develop robust recognition algorithms.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. CS-HMM THEORY

A CS-HMM speech recogniser will resemble a discrete state system in maintaining a list of hypotheses corresponding to phonetic assumptions. Hypotheses specify whether they are in a dwell or a transition, how many time steps they have spent in the current dwell or transition, the identity of the current phoneme (for dwells) or the phoneme recently left (for transitions) and enough phonetic history to apply a language model. The difference is that in a CS-HMM each hypothesis is characterised by a discrete component, here the phonetic history, and by additional continuous state variables. By using a parametric representation, each hypothesis stores information about an infinite set of states.

The information takes the form of a Baum-Welch alpha value, written $\alpha_t(\boldsymbol{x})$ where $\boldsymbol{x} \in \mathbb{R}^d$ is a vector of continuous state variables and t is time. This value is the sum of path probabilities over all paths arriving in state $\boldsymbol{x}$ at time t, where a path probability is the product over previous times of the state probability (conditioned on its predecessor) and the observation probability, and where we loosely use the term 'probability' to denote the value of a probability density function. The paths are limited to those through the same sequence of discrete state components, that is phonetic history.

The alphas are assumed to take the parametric form

$$\alpha_t(\boldsymbol{x}) = K_t\, n(\boldsymbol{x} - \boldsymbol{\mu}_t, P_t) \tag{1}$$

for some scale factor K_t and additional parameters $\boldsymbol{\mu}_t$ and P_t, where for convenience we define

$$n(\boldsymbol{x}, P) = (2\pi)^{-d/2}|P|^{1/2} \exp\{-\tfrac{1}{2}\boldsymbol{x}^T P \boldsymbol{x}\} \tag{2}$$

with P a precision matrix (the inverse of a covariance). The scale factor K_t is the sum of probabilities of all paths consistent with a given hypothesis, which is the right quantity to threshold on when pruning a list of hypotheses. K_t is taken as the score of a particular hypothesis.

We assume that a phoneme inventory, of size N_ϕ, has been created. For each phoneme ϕ we record the canonical formant frequencies $\boldsymbol{f}_\phi$ and precision A with which these frequencies are realised in any particular example of the phoneme. In this work we assume a global A across all phonemes, but this could vary by phoneme. In this work, we refer to formants but the method can be applied to any low-dimensional representation of speech.

We begin the induction by assuming that the first time step is the start of a dwell period for some phoneme ϕ. The continuous state component $\boldsymbol{x}$ is the 3-long vector of realised formant frequency targets whose prior probability density function (PDF) determines the initial alphas

$$\alpha_0(\boldsymbol{x}) = n\big(\boldsymbol{x} - \boldsymbol{f}_\phi, A\big) \tag{3}$$

where $\boldsymbol{f}_\phi$ is from the phoneme inventory. There is an initial hypothesis for each phoneme in the inventory. The initial alphas are written in the form (1), and the inductive calculation shows that this form is retained through the entire time sequence.

Suppose a dwell state at time $t-1$, with the hypothesis parametrised as (1), and observation $\boldsymbol{y}_t$ is made. Assuming Gaussian measurement errors, the observation is drawn from the distribution with PDF $n(\boldsymbol{y}_t - \boldsymbol{x}, E)$ where E is the measurement precision. The hypothesis can be updated to take account of this observation

$$\begin{aligned}\alpha_t(\boldsymbol{x}) &= K_{t-1}\, n\big(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}, P_{t-1}\big)\, n(\boldsymbol{y}_t - \boldsymbol{x}, E)\,, \\ &= K_t\, n(\boldsymbol{x} - \boldsymbol{\mu}_t, P_t)\end{aligned} \tag{4}$$

where

$$P_t = P_{t-1} + E, \tag{5}$$

$$\boldsymbol{\mu}_t = P_t^{-1}\left(P_{t-1}\boldsymbol{\mu}_{t-1} + E\boldsymbol{y}_t\right), \tag{6}$$

$$K_t = K_{t-1}\, n\Big(\boldsymbol{y}_t - \boldsymbol{\mu}_{t-1}, \big(P_{t-1}^{-1} + E^{-1}\big)^{-1}\Big)\,. \tag{7}$$

On entering a transition region, we append 3 formant slope values $\boldsymbol{s}$ to the continuous state component. The alpha for the first step of the transition, given observation $\boldsymbol{y}_t$, is

$$\alpha_t(\boldsymbol{x}, \boldsymbol{s}) = K_t\, n\left(\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{s}\end{pmatrix} - \boldsymbol{\mu}_t, P_t\right) \tag{8}$$

where $\boldsymbol{\mu}_t$ has been extended to a 6-dimensional vector and P_t is a 6×6 precision matrix

$$\boldsymbol{\mu}_t = \begin{pmatrix}\boldsymbol{\mu}_{t-1}\\ \boldsymbol{y}_t - \boldsymbol{\mu}_{t-1}\end{pmatrix}, \tag{9}$$

$$P_t = \begin{pmatrix}P_{t-1} + E & E\\ E & E\end{pmatrix} \tag{10}$$

and $K_t = K_{t-1}$. The observation has been accounted for by setting the slope components appropriately (see (9)) and so the hypothesis score K_t is not changed.

For subsequent steps through a transition region observation $\boldsymbol{y}_t$, made h steps into the transition, is drawn from a distribution with PDF $n(\boldsymbol{y}_t - (\boldsymbol{x} + h\boldsymbol{s}), E)$, and we write the alphas as in (8) with

$$P_t = P_{t-1} + \begin{pmatrix}E & hE\\ hE & h^2E\end{pmatrix}, \tag{11}$$

$$\boldsymbol{\mu}_t = P_t^{-1}\left(P_{t-1}\boldsymbol{\mu}_{t-1} + \begin{pmatrix}E\boldsymbol{y}_t\\ hE\boldsymbol{y}_t\end{pmatrix}\right), \tag{12}$$

$$\begin{aligned}K_t = K_{t-1}&\sqrt{\frac{|P_{t-1}||E|}{|P_t|\,(2\pi)^3}} \\ &\times \exp\{-\tfrac{1}{2}\left(\boldsymbol{\mu}_t^T P_t \boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}^T P_{t-1}\boldsymbol{\mu}_{t-1} - \boldsymbol{y}_t^T E \boldsymbol{y}_t\right)\}.\end{aligned} \tag{13}$$

A more complicated step arises when we move from a transition region to a dwell state. During the transition, hypothesis α_t is parametrised by $\boldsymbol{x}$ and $\boldsymbol{s}$, the frequency values

at the beginning of the transition and the slope in the transition. Assuming the transition has lasted h frames then we wish to reparametrise in terms of $\boldsymbol{x'} = \boldsymbol{x} + h\boldsymbol{s}$ the formant frequencies at the beginning of the next dwell. Thus, we must marginalise against the slope variables

$$\alpha_t(\boldsymbol{x}') = \int \alpha_t(\boldsymbol{x}, \boldsymbol{s})\,\mathrm{d}\boldsymbol{s} = \int \alpha_t(\boldsymbol{x}' - h\boldsymbol{s}, \boldsymbol{s})\,\mathrm{d}\boldsymbol{s}, \quad (14)$$

$$= K_t\, n(\boldsymbol{x}' - \boldsymbol{\mu}', P')\,, \quad (15)$$

with the integrand in (14) being of the form (8). The quantities $\boldsymbol{\mu}'$ and P' must be computed. For full details of this calculation refer to [21].

2.1. Language and phoneme modelling

At the beginning of a dwell region, knowledge of any language model and the expected formant frequencies for each phoneme is included. For simplicity in this paper we use the equivalent of no language model, that is we assume any phoneme may follow any other with equal probability — this is accounted for by the $1/N_\phi$ factor below. Each hypothesis branches into N_ϕ hypotheses, one for each phoneme,

$$\alpha_t^{(i)}(\boldsymbol{x}) = \alpha_t(\boldsymbol{x})\, \tfrac{1}{N_\phi} n(\boldsymbol{x} - \boldsymbol{f}_i, \mathrm{A})\,, \quad (16)$$

where $\alpha_t(\boldsymbol{x})$ taken from (15), dropping the primes.

2.2. Timing models

A second method for branching is given by the timing model, that is the distribution of persistence times for dwell and transition regions. We assume dwell times are integer, drawn uniformly in the range $[d_{\min}, d_{\max}]$ and similarly, transition times are in the range $[t_{\min}, t_{\max}]$.

Suppose a hypothesis has spent h frames in a dwell state with $d_{\min} \le h \le d_{\max}$. Then there is a possibility the next frame will be in a transition region. Therefore, the hypothesis branches with the possibilities being remaining in the dwell, with probability p, or beginning a transition, with probability $1-p$. For a consistent description, the hypothesis likelihood, K_t, is updated with these probabilities. Essentially the same argument applies within transition regions moving to dwells.

2.3. Thresholding

As observations are made, the total number of hypotheses which must be considered grows exponentially. We maintain the best 250 hypotheses provided their likelihoods K_t satisfy $\log K_t > \log\kappa - 100$ where κ is the largest likelihood.

Ideally, we should not make any hard decisions about the phoneme recovery until the end of the audio sample and then the most likely hypothesis holds the recovered sequence of phonemes. For the results shown in this paper, we defer making any decision by 20 frames (0.2s). Any hypothesis which is inconsistent with the (delayed) decision is removed.

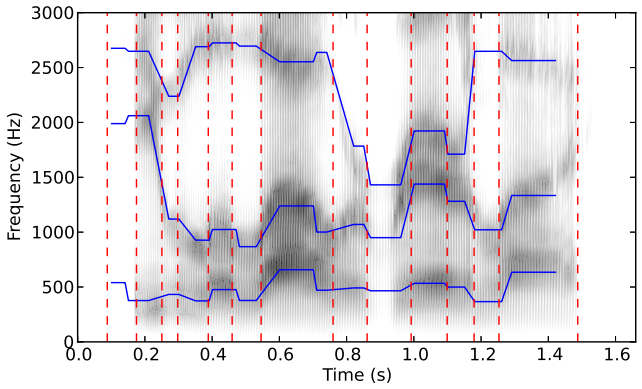

Fig. 1. Spectrogram and CS-HMM recovered formant tracks (blue solid) for TIMIT file test/dr2/mwew0/sx11. Vertical dashed lines (red) show the TIMIT phoneme boundaries.

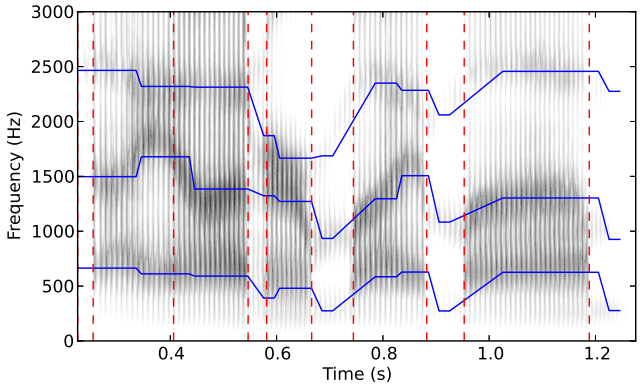

Fig. 2. As Figure 1 but for TIMIT file train/dr6/mrmb0/sx231. The correct transcription is */hh ay aa n er m ay m aa m/* ('I honor my mom'); the CS-HMM finds */hh ay aa n er m aa ay m aa m/*.

3. RESULTS

To illustrate the CS-HMM technique we apply it to formant tracks, using data from the VTR database [22]. This database contains hand corrected formant tracks for a selection of TIMIT sentences [20]. The TIMIT phoneset is reduced to a phoneset of size 39 [23].

Figure 1 shows the results for a single file. Here the correct transcription is */hh iy w l ah l aw er r eh r l ay/* for the phrase 'he will allow a rare lie'. Constructing a phoneme inventory from the VTR data, and assuming the central 70% of a phoneme represents the dwell phase, we recover the phoneme sequence */hh iy w l ah l aw ah er r eh r l ay/*. In most cases the transition regions correspond to the TIMIT phoneme boundaries. These locations have been found as part of CS-HMM recovery. There is a single phoneme inserted, an */ah/* at 0.75s. From the underlying spectrogram, we can see that the transition from */aw/* to */er/* actually has a sharp drop

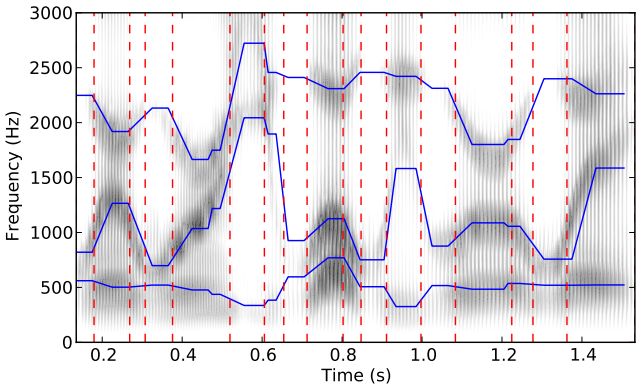

Fig. 3. As Figure 3 but for TIMIT file test/dr8/mjln0/sx9. The correct transcription is */w eh ah w er y uw w aa l w iy w er ah w ey/* ('where were you while we were away'); the CS-HMM finds */w eh w er eh y uw l aa w iy w er ah w ey/*.

in F_2 followed by a rise – this is the cause of the insertion. The */aw/* phoneme is a diphthong. In the CS-HMM recovery we see this diphthong split into component parts. There are two instances of */r/*, one from 0.9–1.0s and the second from 1.1–1.2s. Both are recognised correctly, but the frequency estimates of each are quite different. This demonstrates one strength of the model in explicitly stating that realisations of a particular phoneme come from a distribution.

A second example is shown in Figure 2. Again, there is a single phoneme insertion. The diphthong */ay/* at 0.8s has been split into */aa ay/*. Looking at other occurrences, the system is matching */ay/* as the second part of diphthongs and so this is actually the correct recovery for this file.

Our final example is shown in Figure 3. There are a number of errors, the first being a deleted */ah/* at 0.3s. Here, the VTR data continues in a smooth linear transition and so there is no phoneme dwell as assumed by our model. Listening to the recording it is questionable whether the */ah/* phoneme can be heard — it is labelled in TIMIT as having a duration of less than four frames. The second error is the insertion of */eh/* at 0.45s. This has been inserted as there is a change in the slope of the formant tracks and the only way this can be accomplished, within the model, is to insert a phoneme dwell then begin a second transition. In this region of the audio, F_2 and F_3 become close and so these slope changes could be a result of how the VTR data was generated. The remaining errors are a result of confusion between */l/* and */w/*.

4. DISCUSSION

Our CS-HMM method, similar to Ainsleigh *et al.*'s HGMM [19], models and recovers a speech signal encoded as smoothly varying parameters of speech dynamics. The idea of modelling at the segmental rather than frame level is not new, having been postulated for ASR by Bridle and Ralls [17], and applied in a statistical segmental pattern-matching manner by Paliwal and Rao [18]. Ostendorf *et al.* [8] review various segmental HMM and dynamic modelling approaches.

To the best of our knowledge, this paper is the first to apply a CS-HMM to model speech utterances and recover phoneme sequences, and as such is the most parsimonious speech modelling and recovery algorithm of which we are aware. Previous proposals for models with many fewer parameters than 'standard' discrete HMM and deep neural network approaches (e.g. [3, 4, 14]) have nevertheless required training of orders of magnitude more parameters than the CS-HMM. While our model is in early development and may need extension to account for the full range of characteristics of natural speech, our results give cause for optimism that a generally applicable model can be developed. The full theory will be described in a forthcoming paper [21].

To analyse our approach, we applied it to Deng *et al.*'s vocal tract resonance (VTR) data [22] extracted from TIMIT. We showed that with careful parameter tuning, the CS-HMM can recover sequences of phonemes found in TIMIT utterances, and discrepancies are explainable phonetically or in terms of the algorithm's behaviour. It may be argued that our experimentation is limited and in some ways artificial, but it must be borne in mind that in this early work, we require very few parameters — 3 formant target frequencies for each of 39 phonemes, 4 dwell duration parameters, and 6 non-zero entries in the phoneme target frequency and observation frequency precision matrices. The model carries out the full segmentation and phoneme recovery, rather than re-scoring n-best lists generated by a DS-HMM (e.g. [3, 4]). We use at present no language model.

Our results compare favourably with the results from the early segmental system described by Paliwal and Rao [18], where training and test were on the same set of sentences spoken by a single (male) speaker, on two occasions separated by one week; and also the n-best list rescoring results reported by Deng and Ma [3, 4] when we ignore those results where the reference transcription was included in the n-best list, or where critical phone alignment parameters of the dynamic models were hand-crafted.

5. FUTURE WORK

We have shown that an accurate phoneme inventory is crucial and are currently developing mechanisms for more accurately training this. We are also extending the CS-HMM framework to handle the full range of speech sounds. Two further clear extensions are to improve the timing model, it being reasonably well established that phoneme durations follow approximately a log-normal distribution; and to integrate a language model, which could be quite strong, since a hypothesis may contain many frames of context. We also intend to study more formally the relationships between CS-HMMs, intermediate state HMMs, and segmental models.

6. REFERENCES

[1] J. N. Holmes, "A Parallel-Formant Synthesizer for Machine Voice Output," In Fallside and Woods [24], pp. 163–187, 1985.

[2] K. Iskarous, L. Goldstein, D. H. Whalen, M. Tiede, and P. Rubin, "CASY: The Haskins configurable articulatory synthesizer," in *ICPS, Barcelona*, 2003, pp. 185–188.

[3] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[4] J. Z. Ma and L. Deng, "Target-directed mixture dynamic models for spontaneous speech recognition," *IEEE Trans. Speech Audio Proc.*, vol. 12, no. 1, pp. 47–58, 2004.

[5] S. Young, "A Review of Large-vocabulary Continuous-speech Recognition," *IEEE Signal Process. Mag.*, pp. 45–57, 1996.

[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28 no. 4, pp. 357–366, 1980.

[7] M. J. Russell and R. K. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. ICASSP*, New York, 1985, pp. 5 – 8.

[8] M. Ostendorf and V. Digalakis and O. Kimball, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Proc.*, vol. 4, no. 5, pp. 360–378, 1996.

[9] M. J. F. Gales and S. J. Young, "Segmental hidden Markov models," in *Proc. Eurospeech*. ISCA, 1993.

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

[11] G. E. Henter and W. B. Kleijn, "Intermediate-State HMMs to capture continuously-changing signal features," in *Proc. Interspeech*, Florence, Italy, 2011, vol.12, pp. 1817 – 20.

[12] G. E. Henter, M. R. Frean and W. B. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4505 – 4508.

[13] L. Deng, L. J. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Proc. ICASSP*, 2004, vol. 1, pp. 557–60.

[14] 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," Tech. Rep., Center for Language and Speech Processing, The John Hopkins University, 1998.

[15] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 723–742, 2007.

[16] J. Holmes, I. Mattingly, and J. Shearne, "Speech synthesis by rule," *Language and Speech*, vol. 7, pp. 127–143, 1964.

[17] J. S. Bridle and M. P. Ralls, "An Approach to Speech Recognition using Synthesis-by-Rule," In Fallside and Woods [24], pp. 277–292, 1985.

[18] K. K. Paliwal and P. V. S. Rao, "Synthesis-based recognition of continuous speech," *J. Acoust. Soc. Am.*, vol. 71, no. 4, pp. 1016–1024, 1982.

[19] P. L. Ainsleigh, N. Kehtarnavaz, and R. L. Streit, "Hidden Gauss-Markov models for signal classification," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1355–1367, 2002.

[20] J. S. Garofolo, "TIMIT: Acoustic-phonetic continuous speech corpus," Tech. Rep., Linguistic Data Consortium, 1993.

[21] C. J. Champion and S. M. Houghton, "Application of Continuous State Hidden Markov Models to a Classical Problem in Speech Recognition," Submitted to *Computer Speech and Language*, 2014.

[22] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Yanyi Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. ICASSP*, 2006.

[23] Kai-Fu Lee and Hsiao-Wuen Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[24] F. Fallside and W. A. Woods, Eds., *Computer Speech Processing*, Prentice Hall, Englewood Cliffs, New Jersey, 1985.