社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（1）
2004／12／20

〔招待講演〕

生成モデルを用いた音声認識

マクダーモット・エリック†

† 日本電信電話（株）ＮＴＴコミュニケーション科学基礎研究所　〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: †mcd@cslab.kecl.ntt.co.jp

あらまし　従来の音声認識技術は，音声生成過程の知識をほとんど反映しておらず，観測された表層的な音声特徴のみを扱うものが多い．本稿では，音声生成過程の視点から音声認識を行う取り組みについて紹介する．特に，音声学的特徴の検出を隠れマルコフモデル（HMM）に統合する方法，調音パラメータを記述する連続状態線形動的システム及び音声生成制約を用いたグラフィカルモデル等について論じる．

キーワード　音声認識、音声生成、調音モデル、グラフィカルモデル、線形動的システム

【Invited】

Production models for speech recognition

Erik MCDERMOTT†

† Nippon Telegraph and Telephone Corporation, NTT Communication Science Laboratories 2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto-fu, 619-0237 Japan
E-mail: †mcd@cslab.kecl.ntt.co.jp

Abstract　Acoustic modeling in speech recognition uses very little knowledge of the speech production process. At many levels our models continue to model speech as a *surface* phenomenon. Typically, hidden Markov model (HMM) parameters operate primarily in the acoustic space or in a linear transformation thereof; state-to-state evolution is modeled only crudely, with no explicit relationship between states, such as would be afforded by the use of phonetic features commonly used by linguists to describe speech phenomena, or by the continuity and smoothness of the production parameters governing speech. This article attempts to provide an overview of proposals by several researchers for improving acoustic modeling in these regards. Such topics as the controversial Motor Theory of Speech Perception, work by Hogden explicitly using a continuity constraint in a pseudo-articulatory domain, the Kalman filter based Hidden Dynamic Model, and work by many groups showing the benefits of using articulatory features instead of phones as the underlying units of speech, will be covered.

Key words　speech recognition, speech production, articulatory modeling, linear dynamical systems

1. Introduction

The dominant approach to acoustic modeling in the speech recognition community continues to be the "beads on a string" view of phonetics. The speech signal is essentially modeled as a concatenation of phones. The fact that, due to co-articulation, the acoustic realization of phones is context-dependent is accounted for by the use of triphone or higher-order context models. Since the set of possible contexts is difficult to estimate robustly, statistical clustering techniques must be used. It can be surmised that this approach is not optimal, and perhaps effective only for relatively constrained types of speech situations. Variations in speech production, resulting for example from changes in speaking rate, are currently modeled only by (1) absorption of this variation into the acoustic model parameters, resulting in models with overly broad variances, (2) the creation of situation-specific models, leading again to the problem of robust estimation, or (3) explicit modeling of phonetic re-organization, which is somewhat limited by the coarseness of the phone unit. These limitations can all be viewed as the result of the "beads on a string" approach. Incorporating better models of speech production into ASR may be able to alleviate these problems [1] and as a result significantly improve recognition of spontaneous speech, robustness to noise and the construction of multi-lingual acoustic models.

2. The structure of speech

Speech as articulatory gestures

Many speech recognition engineers would profit from the study of fundamental texts in acoustic and articulatory phonetics [2] [3] [4]. An influential, if controversial, perspective on speech organization is that developed at Haskins Laboratories in the 1980s [5]. A central tenet of this perspective is that speech percepts fundamentally correspond to the articulatory *gestures* that gave rise to the acoustic signal. Gestures typically involve several articulators working together in (loose) synchrony; their description is thus multi-dimensional and time-varying, similar to that of a musical score. An example of such a gesture, defined in terms of the

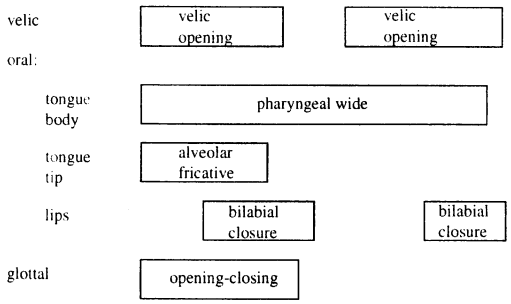

図 1 Gestural score for the word *spam*.

Browman and Goldstein features [6], is shown in Figure 1.

The power of the gestural approach is that it provides a natural account of variation in spontaneous or casual speech. Instead of using complex phonological rewrite rules to account for phenomena such as lenition, reduction and insertion, simple and predictable changes in the temporal relations between different vocal tract variables can account for the same phenomena [6]. A vivid example of the representational power of the gestural approach is provided by [7] for the utterances /banana/, /bandana/, /badnana/ and /baddata/. The differences between these utterances all come down to differences in the timing of velar movement.

3. The Motor Theory

The mechanism by which listeners analyze the acoustic speech signal into articulatory gestures has been the subject of great conjecture and controversy. The *Motor Theory* of speech perception [8] holds that during speech *perception* listeners access the parts of their brains involved in speech *production*.

fMRI-observed neural activity during speech perception

The "strong" version of this theory holds that speech is always perceived using production-related neural mechanisms. Though few still ascribe to this belief, weaker versions of the theory are still influential. It is interesting to note that recent work using fMRI to measure brain activity during speech perception has found evidence that production centers are in fact activated during perception. This happens especially during difficult perceptual tasks, such as perception in noise, or second language perception, but also during normal perception as well [9] [10] [11].

4. Feature detection; integration with existing HMM/hybrid architectures

A number of studies exist in which articulatory (or more generally, phonetic) features are individually estimated, typically with an Artificial Neural Network (ANN), and then used to either replace or augment acoustic observations in an existing HMM system.

4.1 Multi-stream architecture

Metze and Waibel describe a multi-stream architecture in which a standard CD-HMM system is supplemented with feature-detecting Gaussian mixture models (GMMs) [12]. Each of the additional GMMs is a one-layer decision tree representing either the presence or absence of a particular feature. These are trained independently of one another. Several methods were considered for selecting which feature GMMs to combine with the baseline system, such as the selection of feature GMMs according to their feature classification rate, or successively considering overall performance resulting from the addition of each new feature. The best of the methods improved the baseline performance of 13.4% to 11.6% WER on a read Broadcast News task, and also yielded a clear benefit on ESST (Verbmobil) data. It is interesting to note that the features they use come directly from the phonetic decision trees already used by the (standard) context-dependent models of the baseline system.

4.2 Using estimated features to augment the acoustic feature vector

King and Taylor [13] [14] describe an approach in which phonetic features are first estimated from the acoustics using an ANN. The output of the ANN is then used to form a feature vector, that is then passed to a conventional HMM. This study considered both a binary, "distinctive" feature set [15] and the multi-valued feature set used in [16] (similar, it appears, to the Browman and Goldstein features [6]). It was found that replacing the MFCC feature vector with a feature vector based on automatically determined phonetic features yielded nearly identical performance to the original MFCC-based system on the TIMIT task.

Eide [17] describes the use of GMMs to model distinctive features, followed by the use of a mutual information criterion to select the features to use in augmenting the acoustic feature vector.

Kirchhoff [18] [19] [20] describes a similar use of multi-valued feature estimation and integration, but in addition considers both additive and multiplicative feature integration. This work also considers the use of two distinct acoustic representations, PLP/RASTA as well as modulation spectrogram. Acoustic-only, articulatory-feature only, and combined acoustic/articulatory models were evaluated on the OGI Numbers and Verbmobil task, showing significant benefits to acoustic/articulatory combination. Furthermore, benefits of articulatory representations were found in noisy and reverberant environments.

In earlier work [16], Kirchhoff considered a syllable-based parallel feature decoding architecture that used dynamic programming based on string edit distance to find optimal synchronization of multi-featured syllable templates to ANN-detected features. This approach specifically accounts for the well-known phenomenon of partially de-synchronized articulator movement.

Recently, Webster described a different use of ANN-based feature detector with syllable templates, evaluated on the TIMIT task [21].

5. Articulatory configurations as HMM states

5.1 Li Deng and colleagues

An early and well-known approach to introducing articulatory knowledge into HMM systems is that of Deng et al. [22] [23] [24]. In this approach, multi-valued features similar to the Browman and Goldstein features are used to specify different lexical entries in terms of state networks representing different possible feature trajectories. Each HMM state represents a different articulatory feature configuration. The approach is to take the canonical feature representations for each phoneme in a lexical entry, but to then model possible variations in the feature transitions, such as anticipatory or inertial feature spreading. This is a knowledge-oriented way of designing context-dependent states, that explicitly allows for de-synchronization of feature movements, within bounds. A danger with the approach is that the state networks risk becoming very large if too much leeway is allowed for feature de-synchronization. For more recent work with this approach, see [25].

5.2 Related work

A similar approach was investigated by Richardson et al. [26]. Static constraints on allowable articulatory configurations, and dynamic constraints imposing continuity and

limiting maximum articulator velocity, were used to reduce the size of the state network significantly. The approach was shown to provide benefits when tested on noisy speech, and, in combination with a standard HMM, when tested on clean speech as well. Good results were obtained on the large vocabulary isolated word PHONEBOOK task.

6. Between feature bundles and feature detectors: Factorial HMMs

Representing different feature configurations with HMM states is clearly not a parsimonious representation of the data. Extracting individual features from the data independently offers a much more compact representation, but raises questions about feature combination. One approach that attempts to find a happy medium is that of Factorial HMMs, also referred to as "loosely coupled HMMs" [27] [28]. This is a way of representing different streams while modeling varying degrees of coupling between the streams. The architecture is well-suited to the multi-band approach to acoustic modeling, as well as to articulatory-based modeling. However, to our knowledge, no studies have reported the use of factorial HMMs with explicitly production-oriented representations – with the exception of work in audio-visual speech recognition [29] [30]. (Visual observations, for example of lip motion, correspond to direct observation of speech production parameters).

7. Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) [31] are a new approach for statistical modeling that has been applied to speech recognition [32] [33]. Compared to standard HMMs, DBNs may offer a substantially better platform to model production-related effects in speech. In particular, rather than combining hand-coded rules and probabilistic inference – a description that seems to apply to the studies reviewed here using articulatory features in various schemes – DBNs allow a more natural integration of production knowledge with statistical pattern recognition [33]. In principle, dependencies on speaker type and speaking rate, for instance, or on higher-level prosodic structure [34], can be represented more easily than in standard HMMs.

A number of studies have specifically examined the incorporation of articulatory knowledge into DBN structures; in particular, see work by Markov et al. [35] [36], and Livescu et al. [37] [38] [39].

8. Dynamical systems

Several studies have made use of the classic Linear Dynamical System / Kalman filter paradigm:

$$\begin{aligned} \mathbf{x}_t &= A\mathbf{x}_{t-1} + \mathbf{b}_t \\ \mathbf{y}_t &= C\mathbf{x}_t + \mathbf{d}_t, \end{aligned} \qquad (1)$$

where $\mathbf{x}_t$ is the state of the system at time t, $\mathbf{y}_t$ is the observation at time t. A and C are matrices, and $\mathbf{b}_t$ and $\mathbf{d}_t$ are (typically Gaussian) noise. This type of dynamical system is often referred to as a continuous state HMM. In a well-known study, Digalakis et al. proposed the use of a linear dynamical system for speech recognition, and obtained good phoneme classification rates on the TIMIT task [40].

A natural interpretation of this model in the context of speech modeling is that $\mathbf{x}_t$ represents an articulatory or pseudo-articulatory state vector, and $\mathbf{y}_t$ represents the acoustic consequence of that state. Such a model of speech production was investigated in Honda's 1977 doctoral thesis [41]. The use of linear equations to describe both state evolution and mapping from state to observation may be questioned in the case of speech, but linear models can still afford insight [42], and non-linear extensions of the paradigm abound.

8.1 Co-articulation modeling with targets

The use of a spatial target in the articulatory domain to account for co-articulation has attracted many researchers. On this view, co-articulation is the natural result of smooth motion from target to target. In the field of speech production, many ideas have been proposed to achieve this [43]. The targets may be attractors [44], via points [45], spatio-temporal regions [46], or abstract goals which may be defined in articulatory, acoustic, or oro-sensory terms [47].

Bakis proposed the use of targets in a continuous state HMM to model co-articulation for speech recognition [48]. More recently, Richards and Bridle proposed the well-known Hidden Dynamic Model (HDM) [49]. In this approach, low-dimensional phoneme-dependent target vectors are smoothed to yield trajectories in the hidden dynamic space, namely, the state evolution part of Equ. (1). The mapping from hidden state to acoustics is performed with an ANN. The entire architecture is optimized jointly; the phoneme-dependent targets are learned as well. Preliminary evaluation results on Switchboard using the N-best rescoring approach were reported in [50]. An efficient search algorithm for a related approach was investigated in [51].

In most HDM studies, the hidden state space is taken to correspond to formant tracks / vocal tract resonances (VTRs) [52] [53]. Gao et al. [54] proposed to initialize the HDM target vectors using articulatory features derived from phonetic knowledge. They describe use of the same model for both recognition and synthesis.

For an earlier approach to a production-oriented dynamical system, also using ANNs to map from pseudo-articulator positions to acoustics, see work by Blackburn [55] [56].

8.2 Switching State-Space Models

The HDM (again, using target vectors corresponding to vocal tract resonances) has been extended to use the recently proposed Switching State-space Model [57] [58]. This enables the carrying over of the posterior distribution of the state vector across segment boundaries - potentially a crucial issue in modeling co-articulation with linear dynamical systems. Other recent work using switching state-space models for speech recognition includes [59], [60], and [61].

8.3 Use of articulatory data

A number of studies have investigated the use of actual articulatory data, measured for example using Electromagnetic Articulograph (EMA) sensors, to supplement acoustic observations in novel model structures. In particular, King and Wrench [62] describe the use of a linear dynamical system to model various combinations of articulatory data, ANN-estimated articulatory data, acoustic data, and hidden variables in the two layers of the linear dynamical system in Equ. (1). This study also considered different types of parameter tying in Equ. (1). A natural approach might be to tie C across all phone/syllable models, but keep A model-dependent, thereby distinguishing models by their state evolution, yet using a common mapping from state to observations; however, this did not yield the best result. Good syllable classification results were obtained for some of the combinations examined on a small speaker-dependent task. Further work along this direction was reported in [63], which discussed the use of the linear dynamic model with a stack decoder for recognition, and examined articulator criticality in terms of estimated variance. This study used articulatory and acoustic data from the MOCHA database [64]. Sun et al. recently described the use of articulatory data in the context of the Deng-style articulatory-feature based HMMs [65]. Also see work by Blackburn [66].

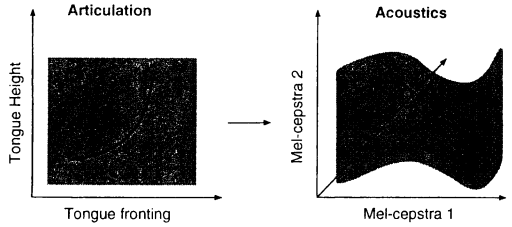

図 2 Manifolds of speech

9. Acoustic-to-articulatory inversion

9.1 Inversion in general

Estimating vocal tract shape and trajectory from acoustic data is one of the classic problems of speech science. Some fundamental difficulties of the inversion problem are discussed in [67]. For recent work using a detailed acoustic and physiological model to estimate vocal tract shape trajectories, see Dang [68], and for a discussion of issues in using inversion for speech recognition, see Bailly [69].

9.2 MALCOM

The Maximum Likelihood Continuity Map (MALCOM) is an original approach to exploiting speech production knowledge for recognition purposes, proposed by Hogden [70]. This is a method for acoustic-to-articulatory inversion that has as its central assumption the fact that articulator trajectories are band-limited. A cutoff frequency of 8-15 Hz is frequently cited; such low-pass filtering is common in the speech production research community [71]. Figure 2 illustrates the scenario. When considered only in acoustic terms, speech utterances can follow highly jagged or discontinuous paths in a high-dimensional acoustic space. However, viewed in articulatory terms, the trajectories are smooth and low-dimensional. Since the acoustic and articulatory spaces are linked by the physics of speech production, this suggests that in fact acoustic speech trajectories are constrained to lie on a manifold determined by the smooth articulatory trajectories and the articulatory-to-acoustic mapping. The idea in MALCOM is to incorporate the smoothness constraint explicitly into a statistical model of speech.

MALCOM uses two layers of representation: an acoustic space, modeled using vector quantization (VQ), and a (low-dimensional) pseudo-articulatory space, modeled with Gaussian pdfs. There is a one-to-one link between VQ codes in the acoustic space and Gaussian pdfs in the pseudo-articulatory space. Incoming (acoustic) speech is VQ encoded; MALCOM then finds a smooth trajectory in the pseudo-articulatory space that maximizes likelihood given the VQ code sequence. This is done elegantly by representing pseudo-articulator motion in the frequency domain. The training phase (in which Gaussian pdfs are estimated in the pseudo-articulatory space) incorporates the smoothness constraint as well.

The recovered pseudo-articulatory positions were compared against real articulatory data – while allowing for an estimate of the rotation, scaling and translation of pseudo-articulatory positions – and good correlation was found. In particular, significantly better correlation was found when using a low cutoff frequency (8-15 Hz) compared to when not band-limiting pseudo-articulator motion [70]. (For a related use of smoothness constraints in acoustic-to-articulatory inversion, see [71]). MALCOM is an unsupervised method that can be applied to acoustic-to-articulatory inversion, but has been extended for recognition as well [72]. Recently the mathematical aspects of this approach to general function inversion have been investigated in greater depth [73].

10. HMMs as production models

Finally, it should be remembered that the HMM, in its conventional form, is itself a production model. With no special measures, HMMs run in production mode yield notoriously poor, barely intelligible speech. However, in recent years, the fact that state-of-the art HMM systems use dynamic acoustic features (e.g. delta and delta-delta MFCC components) has been used to constrain the corresponding synthesized output [74]. The resulting smooth trajectories result in much better speech quality.

10.1 Speech-producing HMMs for speech recognition

Minami et al. have proposed turning the Tokuda approach around, back onto the recognition task [75]. The idea is to use a standard HMM for a first recognition pass. The N-best recognition candidates are then used in the Tokuda synthesis method, i.e., using the delta and delta-delta components to constrain the search for a likelihood-maximizing feature trajectory. The resulting smooth trajectory (and corresponding sequence of model variances) is then used to re-score each recognition candidate. This approach applies the smoothness constraint purely in the acoustic domain, but can be viewed as a production-oriented recognition method. A modified HMM learning method using this approach has also been proposed [76].

10.2 Listening to our acoustic models

The same approach provides an interesting diagnostic for speech recognition research. Since the same system can be used to perform recognition and synthesis [77] [78], one can now listen to the synthesized speech corresponding to a recognition error. Doing so reveals that the Gaussian pdfs used by incorrectly recognized phone sequences to model the utterance can include sounds rather different from their intended coverage. Such analysis by synthesis could be used to identify acoustic modeling problems.

11. Conclusion

Incorporating knowledge about the speech production process into speech recognition systems has attracted the interest of many researchers. Of the approaches considered, the ones that seem to have yielded the best results for practical speech recognition appear to be those that do not require a significant modification of existing HMM architectures. It is interesting to note the parallel between the many studies described using articulatory feature detection and the approach called for by C.-H. Lee [79]. Future work needs to address the practical usability of the more drastic departures from current modeling architectures. Furthermore, no one (to my knowledge) has evaluated speaking-rate conditioning, speaker adaptation, or conditioning on higher-level prosodic structure (such as advocated by Ostendorf [34]), in terms of production-oriented modeling. Finally, only a few studies to date have examined discriminative training of production-oriented models [80]. Such topics should provide rich areas for future research.

文 献

[1] R. C. Rose, J. Schroeter, and M. M. Sondhi, "An investigation of the potential role of speech production models in automatic speech recognition," in *International Conference on Spoken Language Processing*, 1994, pp. 575–578.

[2] P. Ladefoged, *A Course in Phonetics*, 3rd ed. Harcourt Brace, 1993.

[3] G. Fant, *Acoustic theory of speech production*. Mouton, The Hague, 1960.

[4] K. Stevens, *Acoustic Phonetics*. The MIT Press, 1998.

[5] C. Fowler and L. Rosenblum, "The perception of phonetic gestures," in *Modularity and the Motor Theory of Speech Perception*, I. Mattingly and M. Studdert-Kennedy, Eds. Lawren Erlbaum Associates, 1991, ch. 3.

[6] C. Browman and L. Goldstein, "Gestural structures: distinctiveness, phonological processes, and historical change," in *Modularity and the Motor Theory of Speech Perception*, I. Mattingly and M. Studdert-Kennedy, Eds. Lawren Erlbaum Associates, 1991, ch. 13.

[7] P. Rubin and E. Vatikiotis-Bateson, "Measuring and modeling speech production in humans," in *Animal Acoustic Communication: Recent Technical Advances*, S. L. Hopp and C. S. Evans, Eds. New York: Springer-Verlag, 1998, pp. 251–290.

[8] A. M. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[9] D. Callan, A. Callan, K. Honda, and S. Masaki, "Single-sweep EEG analysis of neural processes underlying perception and production of vowels," *Cognitive Brain Research*, vol. 10, pp. 173–176, 2000.

[10] D. Callan, J. Jones, K. Munhall, A. Callan, C. Kroos, and E. Vatikiotis-Bateson, "Neural processes underlying perceptual enhancement by visual speech gestures," *Journal of Cognitive Neuroscience and Neuropsychology*, vol. 14, no. 17, pp. 2213–2218, 2003.

[11] D. Callan, K. Tajima, A. Callan, R. Kubo, and S. Masaki, "Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast," *NeuroImage*, vol. 19, pp. 113–124, 2003.

[12] F. Metze and A. Waibel, "A flexible stream architecture for asr using articulatory features," in *International Conference on Spoken Language Processing*, 2002, pp. 2133–2136.

[13] S. King, T. Stephenson, S. Isard, P. Taylor, and A. Strachan, "Speech recognition via phonetically featured syllables," in *International Conference on Spoken Language Processing*, 1998.

[14] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, pp. 333–345, 2000.

[15] N. Chomsky and M. Halle, *The sound pattern of english*. MIT Press, 1968.

[16] K. Kirchhoff, "Syllable-level desynchronization of phonetic features for speech recognition," in *International Conference on Spoken Language Processing*, 1996.

[17] E. Eide, "Distinctive Features For Use in an Automatic Speech Recognition System," in *Proc. Eurospeech*, 2001.

[18] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *International Conference on Spoken Language Processing*, 1998.

[19] ——, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, Germany, 1999.

[20] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319. 2002.

[21] M. Webster, "Syllable classification using articulatory-acoustic features," in *Proc. Eurospeech*, 2003.

[22] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *Journal of the Acoustical Society of America*. vol. 95, 1994.

[23] L. Deng and K. Erler, "Structural design of hidden Markov model speech recognizer using multivalued phonetic features: comparison with segmental speech units," *Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3058–3067, December 1992.

[24] L. Deng and D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds," in *Proc. IEEE ICASSP*, vol. 1, 1994, pp. 45–48.

[25] D. Sun and L. Deng, "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition," *Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1086–1101, 2002.

[26] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator markov models: performance improvements and robustness to noise," in *International Conference on Spoken Language Processing*, Beijing, China, 1998.

[27] B. Logan and P. Moreno, "Factorial hmms for acoustic modeling," in *Proc. IEEE ICASSP*, 1998, pp. 813–816.

[28] H. J. Nock and S. J. Young, "Loosely coupled hmms for asr," in *International Conference on Spoken Language Processing*, Beijing, China, 1998.

[29] A. Nefian, L. Liang, P. Xiabo, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE ICASSP*, vol. 2, 2002, pp. 2013–2016.

[30] S. Gurbuz, "Robust and efficient techniques for audio-visual speech recognition," Ph.D. dissertation, Clemson University, Dept. of Electrical Engineering, 2002.

[31] K. Murphy, "A brief introduction to graphical models," Institute of Phonetics, University of Saarland - www.ai.mit.edu/ murphyk/Bayes/bayes.html, Tech. Rep., 1998.

[32] J. Bilmes, "Natural Statistical Models for Automatic Speech Recognition," Ph.D. dissertation, University of California, Berkeley, Dept. of EECS, CS division, 1999.

[33] G. G. Zweig, "Speech recognition with Dynamic Bayesian Networks," Ph.D. dissertation, University of California, Berkeley, Computer Science, 2002.

[34] M. Ostendorf, "Incorporating linguistic theories of pronunciation variation into speech-recognition models," *Phil. Trans. R. Soc. Lond.*, vol. 358, pp. 1325–1338, 2000.

[35] K. Markov, J. Dang, Y. Iizuka, and S. Nakamura, "Hybrid HMM/BN ASR System Integrating Spectrum and Articulatory Features," in *Proc. Eurospeech*, 2003, pp. 965–968.

[36] K. Markov, S. Nakamura, and J. Dang, "Integration of articulatory dynamic parameters in HMM/BN based speech recognition system," in *International Conference on Spoken Language Processing*, 2004.

[37] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using Dynamic Bayesian Networks," in *Proc. Eurospeech*, September 2003.

[38] K. Livescu and J. Glass, "Feature-based pronunciation modeling for speech recognition," in *Proc. HLT/NAACL*, May 2004.

[39] ——, "Feature-based pronunciation modeling with trainable asynchrony probabilities," in *International Conference on Spoken Language Processing*, October 2004.

[40] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "A dynamical system approach to continuous speech recognition," in *Proc. ICASSP '91*, Toronto, Canada, May 1991, pp. 289–292.

[41] M. Honda, "Speech feature extraction based on articulatory modeling," Ph.D. dissertation, Waseda University, Department of Science and Engineering, 1977.

[42] M. Russell and P. Jackson, "The effect of an intermediate articulatory layer on the performance of a segmental hmm," in *Proc. Eurospeech*, 2003, pp. 2737–2740.

[43] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 22(2), pp. 93–111, 1997.

[44] E. Saltzman and K. Munhall, "A dynamical approach to

gestural patterning in speech production," *"Ecological Psychology"*, vol. 1, no. 4, pp. 333–382, 1989.

[45] E. Vatikiotis-Bateson, M. Tiede, Y. Wada, V. Gracco, and M. Kawato, "Phoneme extraction using via point estimation of real speech," in *International Conference on Spoken Language Processing*, 1994, pp. 631–634.

[46] S. Suzuki, T. Okadome, and M. Honda, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," in *International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[47] J. S. Perkell, M. L. Matthies, M. A. Svirsky, and M. I. Jordan, "Goal-based speech motor control: a theoretical framework and some preliminary data," *Journal of Phonetics*, vol. 23, pp. 23–35, 1995.

[48] R. Bakis, "Coarticulation modeling with continuous state hmms," in *Proc. IEEE Workshop Automatic Speech Recognition*, Arden House, New York, 1991, pp. 20–21.

[49] H. Richards and J. Bridle, "The HDM: a segmental Hidden Dynamic Model of coarticulation," in *Proc. IEEE ICASSP*, 1999.

[50] J. Picone, S. Pike, R. Regan, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster, "Initial evaluation of Hidden Dynamic Models on Conversational Speech," in *Proc. IEEE ICASSP*, 1999.

[51] J. Ma and L. Deng, "Optimization of dynamic regimes in a statistical hidden dynamic model for conversational speech recognition," in *Proc. Eurospeech*, vol. 3, 1999, pp. 1339–1342.

[52] L. Deng, I. Bazzi, and A. Acero, "Tracking vocal tract resonances using an analytical non-linear predictor and a target-guided temporal constraint," in *Proc. Eurospeech*, vol. 1, 2003, pp. 73–76.

[53] 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. IEEE ICASSP*, vol. 1, 2004, pp. 557–560.

[54] Y. Gao, R. Bakis, J. Huang, and B. Xiang, "Multistage coarticulation model combining articulatory, formant and cepstral features," in *International Conference on Spoken Language Processing*, Beijing, China, 2000.

[55] C. Blackburn and S. Young, "Towards improved speech recognition using a speech production model," in *Proc. Eurospeech*, 1995, pp. 1623–1626.

[56] C. Blackburn, "Articulatory Methods for Speech Production and Recognition," Ph.D. dissertation, Cambridge University, Engineering Department, 1996.

[57] Z. Ghahramani and G. Hinton, "Variational Learning for Switching State-Space Models," *Neural Computation*, vol. 12, no. 4, pp. 831–864, April 2000.

[58] L. Lee, H. Attias, and L. Deng, "Variational Inference And Learning For Segmental Switching State Space Models of Hidden Speech Dynamics," in *Proc. IEEE ICASSP*, vol. 1, 1990, pp. 357–360.

[59] A. Rosti and M. Gales, "Rao-Blackwellised Gibbs sampling for switching linear dynamical systems," in *Proc. IEEE ICASSP*, vol. 1, 2004, pp. 809–812.

[60] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proc. IEEE ICASSP*, vol. 1, 2004, pp. 953–956.

[61] L. Lee, H. Attias, L. Deng, and P. Fieguth, "A multimodal variational approach to learning and inference in switching state space models," in *Proc. IEEE ICASSP*, vol. 5, 2004, pp. 505–508.

[62] S. King and A. Wrench, "Dynamical system modelling of articulator movement," in *International Congress on Phonetic Sciences*, 1999, pp. 2259–2262.

[63] J. Frankel and S. King, "ASR - Articulatory Speech Recognition," in *Proc. Eurospeech*, 2001.

[64] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th seminar on speech production: models and data*, 2000.

[65] J. Sun, L. Deng, and X. Jing, "Data-driven model construction for continuous speech recognition using overlapping articulatory features," in *International Conference on Spoken Language Processing*, vol. 1, 2000, pp. 437–440.

[66] C. S. Blackburn and S. J. Young, "Pseudo-articulatory speech synthesis for recognition using automatic feature extraction from X-ray data," in *International Conference on Spoken Language Processing*, vol. 2, Philadelphia, PA, 1996, pp. 969–972.

[67] B. Atal, J. Chang, M. Mathews, and J. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *IEEE Transactions on Computers*, vol. 63, pp. 1535–1556, 1978.

[68] J. Dang and K. Honda, "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," *Journal of Phonetics*, vol. 30, pp. 511–532, 2002.

[69] G. Bailly, C. Abry, L.-J. Boe, R. Laboissiere, P. Perrier, and J.-L. Schwartz, "Inversion and speech recognition," in *Proc. of EUSIPCO-92*, vol. 1, 1992, pp. 159–164.

[70] J. Hogden, "A maximum likelihood approach to estimating speech articulator positions from speech acoustics," *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2663–2664, October 1996.

[71] H. Yehia, "A study on the speech acoustic-to-articulatory mapping using morphological constraints," Ph.D. dissertation, Nagoya University, Graduate School of Engineering, 2002.

[72] J. Hogden and P. Valdez, "Bridging the gap between speech production and speech recognition," in *5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Germany, 2000.

[73] J. Hogden, P. Valdez, S. Katagiri, and E. McDermott, "Blind inversion of multidimensional functions for speech enhancement," in *Proc. Eurospeech*, 2003, pp. 1409–1412.

[74] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," in *Proc. IEEE ICASSP*, vol. 2, 1999, pp. 585–588.

[75] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "A recognition method with parametric trajectory synthesized using direct relations between static and dynamic feature vector time series," in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. 957–960.

[76] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on hmms with the explicit relationship between static and dynamic features," in *Proc. Eurospeech*, 2003, pp. 865–868.

[77] T. Irino, Y. Minami, T. Nakatani, M. Tsuzaki, and H. Tagawa, "Evaluation of a speech recogniton / generation method based on HMM and STRAIGHT," in *International Conference on Spoken Language Processing*, 2002, pp. 2545–2548.

[78] K. Stevens, "Toward a model for speech recognition," in *Journal of the Acoustical Society of America*, January 1960, pp. 47–55.

[79] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition," in *International Conference on Spoken Language Processing*, 2004.

[80] E. McDermott and T. J. Hazen, "Minimum Classification Error training of landmark models for real-time continuous speech recognition," in *Proc. IEEE ICASSP*, vol. 1, 2004, pp. 937–940.