MOVING BEYOND THE ‘BEADS-ON-A-STRING’ MODEL OF SPEECH

M. Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, WA 98195

ABSTRACT

The notion that a word is composed of a sequence of phone segments, sometimes referred to as ‘beads on a string’, has formed the basis of most speech recognition work for over 15 years. However, as more researchers tackle spontaneous speech recognition tasks, that view is being called into question. This paper raises problems with the phoneme as the basic subword unit in speech recognition, suggesting that finer-grained control is needed to capture the sort of pronunciation variability observed in spontaneous speech. We offer two different alternatives – automatically derived subword units and linguistically motivated distinctive feature systems – and discuss current work in these directions. In addition, we look at problems that arise in acoustic modeling when trying to incorporate higher-level structure with these two strategies.

1. INTRODUCTION

It has often been noted that automatic speech recognition performance is much worse on spontaneous speech than on carefully planned or read speech. For the best systems reporting results on the 1999 DARPA Broadcast News benchmark tests, error rates on the spontaneous speech portion of the test set (14-16%) were nearly double those on the baseline condition of planned, studio recordings (8-9%) [1]. Those sites that also participated in a workshop on conversational speech recognition a few months later reported word error rates of roughly 40%. Pronunciation variability has frequently been cited as a key reason for the poor performance; yet, phone-based pronunciation modeling work has so far led to only small error rate reduction. Could it be that the reliance on the idea of words as a sequence of phoneme segments (‘beads on a string’) has had its day?

In this paper, we will look at evidence against the phoneme as a basic unit in speech recognition and at two alternative lexical representations: automatically derived (sub-phone) units and linguistically motivated states defined in terms of categorical features. In both cases, the goal is finer-level unit control. However, we also acknowledge the need for introducing context dependence on syllable and higher-level structure and discuss mechanisms for doing this. Finally, we discuss the importance of new acoustic modeling research to support the increase in granularity without an explosion of model parameters.

2. THE CASE AGAINST THE PHONEME

Several studies have pointed to acoustic variability as a key problem for systems recognizing spontaneous speech. For example, an SRI study showed near doubling of word error rates on the exact same word sequence when it was spoken spontaneously vs. read [2]. More recently, McAllister *et al*. use simulated data in experiments that suggest that poor pronunciation modeling accounts for the bulk of the high error rate on the Switchboard task [3]. Not surprisingly, there have been a large number of research efforts devoted to pronunciation modeling in the last few years, including techniques that use automatic learning, hand-written phonological rules and various combinations of the two. Unfortunately, the gains from phone-based pronunciation modeling techniques have been disappointing, e.g. reducing word error rates from 40.9% to 38.5% on conversational speech [4]. This gain represents a statistically significant improvement on a difficult task, but not the factor of five reduction predicted in [3]. Of course, the factor of five is optimistic because of the match between modeling assumptions in the recognition and simulation of data, but most researchers still share the intuition that there is more to be gained from pronunciation modeling. Many of the pronunciation models that have been applied are quite sophisticated and work well on read speech, which raises the question: is recognition performance limited by the assumption that pronunciation variation is represented in terms of phone-level substitutions, deletions and insertions?

In an extensive series of experiments with different pronunciation models and training conditions, Saraclar *et al*. show that improving phone recognition accuracy can actually *hurt* word recognition accuracy [5]. Results in [4] may explain this in part: decision-tree pronunciation models generate word-level pronunciation probabilities that do not match the relative frequency of those pronunciations in the data – a flaw in the assumption of conditional independence of phones. (Of course, it is also the case that, in theory, optimizing for accuracy of low-level unit recognition is not the best choice for recognizing higher-level units when the low-level units are sequentially dependent.) The conditional independence assumption can be ameliorated by syllable-level pronunciation prediction, but word error rate reduction is still less than 10% [6].

Another indicator of problems with the phoneme is that phonetic transcription of conversational speech is quite difficult for human labelers. It has been observed, in the Switchboard corpus and in other studies, that phonemes which appear to be ‘deleted’ (in the sense of having little or no identifiable associated time segment in a spectrogram representation) are often still perceived because of the presence of coarticulation effects on neighboring segments. Such short segments are quite frequent, as evidenced by distributional data in hand-labeled phonetic transcriptions [7] and by the high percentage of phones mapped to the minimum allowed duration in a forced alignment using a single-pronunciation dictionary (observed in several studies). In [6], it is noted that the relatively high rate of occurrence of phenomena such as feature spreading and cue trading posed difficulties for labelers transcribing the Switchboard corpus. These phenomena also pose difficulties for phone-based computer recognition models. For example, if a phone is deleted in an alternate pronunciation, a different triphone will be used and coarticulation effects cannot be captured. In fact, this sort of feature spreading may be better captured without explicit phone deletion in the word pronunciation, since the triphone models may have effectively learned the deletion pattern. Note that, in standard HMM training, which is not constrained by hand-labeled phone segment times, triphones learn coarticulation effects that result in 'phonetic' time alignments that do not correspond to where a human labeler would put a phone segment boundary. This behavior of automatically trained triphones is yet another argument against the phone.

Analyses of the hand-labeled Switchboard corpus in terms of deviations from the canonical dictionary pronunciation show a strong dependence on syllable structure, e.g. syllable onsets are most often preserved and codas are most often deleted [7]. For these and other reasons, several researchers have recently argued for the syllable as an alternative to the phoneme for representing speech. In this paper, we take a different tack and argue for finer-grained low-level representation, incorporating dependence on syllable (and higher level) structure via context conditioning. There are several reasons for looking at a finer grained temporal scale. First, using a pronunciation model based on phones but acoustic models based on triphones means that a phone substitution translates into a 3-segment (or, 9-state) substitution which may be an inappropriately long timespan, as pointed out by Saraclar *et al.* [5] who find improved performance using state-level (vs. phone-level) pronunciation modeling. Alternative views of the 'hidden state' of the speech process – either as a vector of articulator trajectories (essentially continuous valued) or as parallel asynchronous streams of binary features – all point to the need for a fine-grained (larger) state space. The need for more temporal detail is also supported experimentally by observations such as improved performance from increasing the number of HMM states per triphone (e.g. [8]) and bigger gains from adding parameters to characterize temporal variability vs. mixture components [9]. Lastly, the need for a state-level generalization mechanism to handle unseen triphones (and syllables) argues for a finer-grained representation.

In the two sections to follow, we will suggest two quite different alternatives – data-driven and linguistically based – for increasing temporal resolution while at the same time retaining a connection to syllable structure.

3. ACOUSTICALLY-DERIVED SUB-WORD UNITS

Acoustically derived sub-word units (ASWUs) represent a data-driven approach to defining the sub-word units of speech. Recognition system design involves a combination of automatic segmentation into stationary regions or 'segments', clustering the segments based on acoustic similarity, and dictionary[1] design. ASWUs were proposed several years ago [10, 11, 12, 13], but they faded from view as speaker-independent recognition became the primary goal, because of the difficulty of distinguishing speaker variability from real pronunciation differences. However, this problem has recently been addressed by integrating the unit and dictionary design step [14, 15], so that an ASWU system is now a viable option for speaker-independent recognition. For read speech tasks and especially for low complexity systems, ASWU HMM systems consistently outperform phone-based systems, giving word error rate reductions of 10-20% for systems of equivalent complexity. Even the limiting requirement of having several instances of each word in the vocabulary can be addressed by using a hybrid phone and ASWU system [16]. The problem of modeling cross-word contextual variations is addressed in [17], and multiple pronunciation dictionary design is covered in [18].

Automatically derived units have the potential for capturing effects associated with syllable and word position, because the assignment of unit sequences to a word pronunciation is completely based on acoustics. However, the connection to syllable structure can be made more explicit by learning ASWU units and pronunciations from syllable tokens rather than word tokens. Using syllable-level tokens would ameliorate the unseen word problem in large vocabulary recognition to some extent, but there will still be many unobserved syllables, particularly with conditioning on lexical stress.

An alternative means of incorporating syllable structure (and modeling state-level pronunciation variation) is to think of ASWU design as essentially the same problem as HMM topology design. One could apply the successive state splitting (SSS) algorithm [19], which has been used for designing triphone state sharing, at the syllable level. The SSS algorithm is essentially a generalization of standard HMM tree-based clustering techniques, e.g. [20], which can learn both contextual and temporal structure (i.e. the topology is not fixed to a certain number of states per phone). Applied at the syllable level, it can easily learn effects of syllable structure. In addition, SSS can incorporate lexical stress and word position by labeling syllables with this information as an extra context conditioning variable that can be used in state splitting. In standard decision tree clustering, this strategy for adding conditioning variables has been referred to as 'tagged clustering,' i.e. phones are tagged with stress and other features and tri-tag (vs. triphone) models are clustered. The idea of tagged clustering was first introduced in speech synthesis by Donovan [21], and subsequent application to recognition has been described in [22, 23, 24]. A limitation of tagged clustering is that coding phones (or syllables) causes a huge increase in the number of elementary context-dependent models, which leads to large memory requirements and increased complexity of training because of the increase in possible data divisions. As a result, only simple tag sets have been explored in large vocabulary systems using cross-word context. Work in progress on multi-stage clustering may address this problem by using different subsets of features in different stages of tree (or topology) design.

Another class of approaches that falls under the data-driven theme is the work on state-level pronunciation modeling, different variations of which have been proposed in [25, 5]. The motivation, as raised in the previous section, is that there are many instances where it is more appropriate to substitute or delete part of a triphone rather than the whole triphone. In this work, the subword units are sequential 'regions' of phones trained using standard triphone design techniques, but the final pronunciation network is not constrained to maintain the original phone-level sequence relationships. While the work reported so far has not taken advantage of syllable structure, it is easy to imagine doing so by starting with triphone states designed using tagged clustering or using decision trees for finding state transformation probabilities.

4. LINGUISTICALLY-MOTIVATED ALTERNATIVES

In linguistics, it is features and not phonemes that are viewed as the fundamental units of speech [26], where phones are specified (or

[1] The term 'dictionary' is used to mean 'pronouncing dictionary', primarily for brevity.

coded) in terms of distinctive features. (Note that the term 'feature' is most often used in the speech recognition literature to refer to acoustic observations, such as cepstral vectors or voice onset time, but here we use 'features' to mean symbolic indicators of phonetic contrasts.) For the most part, distinctive features are related to the manner in which a speech sound is produced (the degree of constriction in the vocal tract), the particular articulator that is used (glottis, soft palate, lips and tongue blade, body and root) and/or place of constriction, and how an articulator is used to produce the sound. Different feature systems have been proposed, including binary and multi-valued features; for simplicity we will restrict our discussion to binary features, with the caveat that feature values can sometimes be unspecified in the 'code' for a phoneme, which could be thought of as a third value. Examples of binary features are nasal, voiced, continuant, labial, etc.

Pronunciation variations can be expressed in terms of context-dependent rules describing changes in the feature values or in feature association with segments. Examples include devoicing of a vowel or final consonant in the context of a subsequent voiceless consonant, reducing a tense vowel 'iy' to a lax 'ih', and changing the place of articulation so that 'n' becomes 'm' when the 'n' is followed by a labial stop (as in 'can be'). Feature changes can also account for apparent phone segment deletion where there is still evidence for the segment in the realization of neighboring segments, as in a nasalized 'ae' in a reduced form of 'can't' or the single dental-nasal segment sometimes produced for the two consonants in 'in the.' Features cannot always be mapped to synchronous parallel time functions, and asynchrony can lead to cases where segments appear to be inserted, as in an epenthetic stop in 'warmth' due to asynchronous changing of the nasal and continuant features.

The goal of a feature-based coding of the HMM state space is to represent such pronunciation variability in terms of asynchronous linguistic feature changes. A word has a lexical representation that is a sequence of d-dimensional symbolic feature vectors, which expands into an asynchronous time sequence, which is mapped to d-dimensional hypercube of states for decoding. In other words, the bit vector that corresponds to the feature values indexes an HMM state, and the state transitions are governed by feature spreading characteristics. The key problems with using the feature representation are simplifying search and estimation of that high dimensional space which, like triphones, will include many states that are never observed.

Deng and colleagues [27, 28] proposed a set of parallel discrete feature streams, with hand-written rules for constraining feature 'spreading.' (Their 'features' correspond to quantized vocal tract shape parameters, but the basic idea applies directly to the distinctive linguistic features discussed here.) The feature vector points to a state model index, and the collection of states defined by the feature spreading rules combine to form what is effectively a context-dependent HMM with state sharing determined by human knowledge rather than automatic clustering. The initial work used independent training of the composite states, which corresponds to assuming that all features are interdependent and has no mechanism for training unseen states. Recent work takes a first step at extending triphone clustering techniques to this paradigm [29], though more research is needed.

A more flexible structure treats the different features and their associated acoustic parameters as independent streams synchronized at the syllable level [30, 31]. By treating the streams as independent, a complex state space is achieved while at the same time keeping the training and decoding problems relatively simple. The framework nicely accommodates a variety of different acoustic measures, which can lead to improved performance in high noise (0dB) conditions [32] and results in reduced confusion between certain phonemes [33]. Decoupling features from phones may also lead to models that generalize better across languages.

The use of completely independent streams may be a bit too flexible, however, as evidenced by the fact that a more traditional phone-based model outperforms the feature-based system in low noise conditions [32]. Two main problems stand out. First, it has been observed that certain sets of features tend to spread or modify together in groups that can be characterized by a hierarchical organization [34]. Thus, the timing of different feature streams needs to be more coordinated, though the existence of the hierarchy facilitates modeling, as proposed in [35]. Secondly, the acoustic correlates of the different features are not strictly independent; there are interactions between some features that enhance certain phonetic contrasts [36]. Such interactions imply that acoustic observation models should be conditioned on sets of features and not individual features. The work of Bilmes on learning model structure [37] may provide an automatic mechanism for learning an appropriate dependence structure that also keeps the model dimensionality small.

5. DYNAMIC PRONUNCIATION MODELS

Once one accepts the role of syllable (and/or word) structure in modeling acoustic variability, which is by now quite clearly established, the question is raised as to whether there might be a role for higher-level structure. Indeed, there appears to be evidence for word frequency, syntax and/or prosodic factors. Fosler-Lussier *et al.* show an interaction between speaking rate and word frequency in predicting how much a word pronunciation will deviate from a dictionary baseform [6]. Syntax appears to be a factor as well – one can say 'gonna' for 'going to' for the infinitive 'to' but not for the preposition.

However, such phenomena may be more directly described in terms of prosodic structure [38], i.e. the perceived emphasis and chunking patterns of speech that are related to (but not identical to) syntactic constituents. Cross-word boundary phonological changes, including 'gonna' but also assimilation as in 'gas shortage,' typically do not occur at major prosodic phrase boundaries, and other insertion-like effects do occur at prosodic boundaries. Dilley *et al.* [39] found that glottalization was more likely at vowel-initial word boundaries when those words were pitch-accented and/or in word-initial position of prosodic phrase boundaries. The frequency of glottalization increased with increased saliency of the location, such that glottalization was quite likely ($> 90\%$ for the female subjects) if a word was both accented and phrase-initial. There may also be an effect of enhanced phonetic realization via 'inserted' features at particularly salient regions of the speech signal. In the Switchboard corpus, there are at least anecdotal examples, e.g. an off-glide of 'ae' is 'enhanced' in an emphasized pronunciation of 'and' resulting in 'ae eh n d' (using a phonetic alphabet). We conjecture that conditioning feature changes on a prosodic hierarchy, starting from the level of the syllable, will be needed to better explain the pronunciation variability in speech.

The dependence of pronunciation variability on higher-level linguistic structure is of great importance to speech recognition systems, because it provides a means of dynamically varying pronunciation probabilities. When all the observed pronunciations of a word are allowed in speech recognition decoding, performance degrades because of the increased confusability between words, e.g. allowing 'ae n' as a pronunciation for 'and' increases the possibility of confusing 'and' and 'an'. For this reason, researchers have begun exploring methods for introducing higher-level structure within the context of the standard statistical (e.g. HMM) recognition paradigm, but taking advantage of multi-pass search architectures to condition on hypothesized word context. But how can high-level structure be incorporated at the same time as the granularity of the model is shrinking?

The answer is really no different than for phone-level modeling. In a multipass search framework, it is possible to condition a word pronunciation model on a broader context, leading to dynamic pronunciation probabilities, as in [22, 40, 6]. The critical, and as yet unanswered question, is at what stage to introduce higher-level context conditioning in unit design. It is impractical to automatically learn structure – whether in terms of acoustically derived units or feature interdependence – when clustering is based on atomic units with only a few (if any) observations. In the data-driven approach, we are currently exploring different alternatives in a multi-stage clustering paradigm.

6. IMPLICATIONS FOR ACOUSTIC MODELING

In this paper, we have raised questions about the phoneme as a suitable sub-word unit for speech recognition and argues for moving to a finer-grained representation. At the same time, we acknowledge that there is a clear dependence on higher level structure that should be accounted for via context conditioning in a dynamic pronunciation model. Alternatives for defining finer-grained units are described based on acoustically-derived or data-driven approaches and linguistically-motivated feature coding of the state space.

In the above discussion, we assumed that the acoustic model is a discrete state HMM, and there are several interesting research paths to pursue within this framework. However, within the discrete state framework, there is a serious problem of explosion of the parameter space, as alluded to earlier. Certainly much can be done in the short term with clever clustering schemes and HMMs will long be relied on in early stages of a multipass search, but the huge number of parameters associated with simple HMM extensions to a large state space calls the approach into question. By Occam's razor, we should be striving for a more parsimonious model. The distinctive feature representation offers the potential for a simplified model if the feature streams are sufficiently independent, but there is evidence that the timing is fairly systematic with respect to higher level structure. Results in robust recognition that argue for multi-rate feature streams further complicate this picture.

The key point that these arguments lead to is that moving away from the 'beads on a string' model is not simply a pronunciation model or unit design issue – it is also an acoustic modeling problem. Changes to the pronunciation model are most likely to succeed if matched with an appropriate acoustic model. Improved acoustic models may require additional layers of hidden states at different time scales, mixed memory Markov models [41], a mixed continuous and discrete hidden state [42], a discrete event model [43], and/or other alternatives. Active research on such alternatives is critical to the advancement of speech recognition.

Acknowledgments

This work was supported by the National Science Foundation, Grant no. ISI-9618926. The views and conclusions contained in this document are those of the author and should not be interpreted as reflecting the official policies of the funding agency.

7. REFERENCES

[1] D. Pallett, J. Fiscuss, J. Garofolo, A. Martin & M. Przybocki, "1998 Broadcast News benchmark test results: English and non-English word error rate performance measures," *Proc. DARPA Broadcast News Workshop,* pp. 5-12, 1999.

[2] M. Weintraub, K. Taussig, K. Hunicke-Smith & A. Snodgrass "Effect of Speaking Style on LVCSR Performance," *Proc. Int. Conf. on Spoken Language Proc.,* supplement, 1996.

[3] D. McAllister, L. Gillick, F. Scattone & M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 1847-1850, 1998.

[4] M. Riley *et al.,* "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* pp. 109-119, 1998.

[5] M. Saraclar, H. Nock & S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models" *Proc. European Conference on Speech Comm. and Tech.,* pp. 515, 1999.

[6] E. Fosler-Lussier, S. Greenberg & N. Morgan, "Incorporating contextual phonetics into automatic speech recognition," *Proc. Int. Congress on Phonetic Sciences,* pp. 611-614, 1999.

[7] S. Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* pp. 47-56, 1998.

[8] J. Billa *et al.,* "Multilingual speech recognition: The 1996 BYBLOS Callhome system," *Proc. European Conference on Speech Comm. and Tech.,* pp. 363-366, 1997.

[9] A. Kannan & M. Ostendorf, "A comparison of trajectory and mixture modeling in segment-based word recognition," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 327-330, 1993.

[10] C.H. Lee, B.-H. Juang, F.K. Soong & L. Rabiner, "Word recognition using whole word and subword models," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 683-686, 1989.

[11] T. Svendsen, K.K. Paliwal, E. Harborg, & P.O. Husøy, "An improved subword based speech recognizer," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 108-111, 1989.

[12] K.K. Paliwal, "Lexicon building methods for an acoustic sub-word based speech recognizer," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 729-732, 1990.

[13] L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer & M.A. Picheny, "A method for the construction of acoustic Markov models for words," *IEEE Trans. Speech and Audio Proc.,* **1**(4) 443-452, 1993.

[14] T. Holter & T. Svendsen, "Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units," *Proc. of the IEEE Workshop on Automatic Speech Recognition,* pp. 199-206, 1997.

[15] M. Bacchiani & M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communications,* to appear.

[16] M. Bacchiani & M. Ostendorf, "Using automatically-derived acoustic subword units in large vocabulary speech recognition," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 1843-1846, 1998.

[17] M. Bacchiani, *Speech recognition system design based on automatically derived units* Boston University Ph.D. dissertation, 1999.

[18] T. Holter & T. Svendsen, "Maximum likelihood modeling of pronunciation variation," *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* pp. 63-66 1998.

[19] M. Ostendorf & H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language,* **11**(1) 17-42, 1997.

[20] S. Young, J. Odell & P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 307-312, 1994.

[21] R. Donovan, *Trainable Speech Synthesis*. University of Cambridge Ph.D. dissertation, 1996.

[22] M. Ostendorf *et al.,* "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," Boston University Technical Report No. ECE-97-002, 1997.

[23] D. Paul, "Extensions to phone-state decision-tree clustering: single tree and tagged clustering," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 1487-1490, 1997.

[24] W. Reichl & W. Chou, "A unified approach of incorporating general features in decision-tree based acoustic modeling," *Proc. Int. Conf. on Acoust., Speech and Signal Proc.,* pp. 573-576, 1999.

[25] E. Eide, "Automatic modeling of pronunciation variations," *Proc. DARPA Broadcast News Workshop,* pp. 95-98, 1999.

[26] M. Halle, "Phonological features," in *International encyclopedia of linguistics,* W. Bright, ed., Oxford University Press 1992.

[27] L. Deng & K. Erler, "Structural design of HMM speech recognizer using multi-valued phonetic features: comparison with segmental speech units," *J. Acoust. Soc. Am.,* **92** 3058-3067, 1992.

[28] K. Erler & G. Freeman, "An HMM-based speech recognizer using overlapping articulatory features," *J. Acoust. Soc. Am.,* **100**(4) 2500-2513, 1996.

[29] L. Deng & J. Wu, "Hierarchical partition of the articulatory state space for overlapping-feature based speech recognition," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 2266-2269, 1996.

[30] K. Kirchhoff, "Syllable-level desynchronisation of phonetic features for speech recognition," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 2274-2276, 1996.

[31] S. King, T. Stephenson, S. Isard, P. Taylor & A. Strachan, "Speech recognition via phonetically featured syllables," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 1031-1034, 1998.

[32] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 891-894, 1998.

[33] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information,* University of Bielefeld, Ph.D. dissertation, 1999.

[34] G. Clements, "The geometry of phonological features," *Phonology Yearbook 2* 223-252, 1985.

[35] M. Ostendorf, "Incorporating linguistic theories of phonological variation into speech recognition models," *RS Phil. Trans.,* to appear.

[36] K. Stevens & S. Keyser, "Primary features and their enhancement in consonants," *Language* **65**(1) 81-106, 1989.

[37] J. Bilmes, *Natural statistical models for automatic speech recognition,* U.C. Berkeley, Dept. of EECS, Ph.D. dissertation, 1999.

[38] S. Shattuck-Hufnagel & A. Turk, "A prosody tutorial for investigators of auditory sentence processing," *J. Psycholinguistic Research,* **25**(2) 193-247, 1996.

[39] L. Dilley, S. Shattuck-Hufnagel & M. Ostendorf, "Glottalization of vowel-initial syllables as a function of prosodic structure," *J. Phonetics,* **24** 423-444, 1996.

[40] M. Finke & A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," *Proc. European Conference on Speech Comm. and Tech.,* pp. 2379-2382, 1997.

[41] L. Saul & M. Jordan, "Mixed memory Markov models: decomposing complex stochastic processes as mixtures of simpler ones," *Machine Learning,* to appear.

[42] M. Ostendorf, V. Digalakis & O. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech and Audio Proc.,* **4**(5) 360-378, 1996.

[43] P. Niyogi, P. Mitra & M. M. Sondhi, "A detection framework for locating phonetic events," *Proc. Int. Conf. on Spoken Language Proc.,* pp. 1067-1070, 1998.