Prosodic Hierarchy as an Organizing Framework for the Sources of Context in Phone-Based and Articulatory-Feature-Based Speech Recognition

Mark Hasegawa-Johnson,[1] Jennifer Cole,[1] Ken Chen,[2]
Partha Lal,[3] Amit Juneja,[4] Taejin Yoon,[1] Sarah Borys,[1] and Xiaodan Zhuang[1]

1. University of Illinois at Urbana-Champaign
2. Washington University
3. University of Edinburgh
4. Think-a-Move, Ltd.

May 8, 2007

Abstract

Automatic speech recognition (ASR) is like solving a crossword puzzle. Context at every level is used to resolve ambiguity: the more context we can bring to bear, the higher will be the accuracy of the ASR. One of the ways in which ASR uses context is by defining context-dependent phonological units. This paper reviews and applies two types of phonological units that we find useful in ASR: "phones" (segmental units), and "articulatory features" (units roughly corresponding to bundles of articulatory gestures and/or quantized tract variables). Although the details of the phone or feature inventory vary from system to system, the requirements for a phone or feature inventory are easy to define: each phone (or each vector of articulatory features) must be both "acoustically compact" (the acoustic correlates of a phone or feature vector are predictable) and "phonologically compact" (the phone or feature correlates of a word, in context, are predictable). This paper proposes that the two goals of a phone inventory may be satisfied by defining phones that are sensitive to prosodic context, or alternatively, by using prosodic context to constrain the temporal evolution of recognized articulatory features. Example systems are described that incorporate contextual constraints from five different levels of the prosodic hierarchy, and from the prosodic disruptions caused by disfluency. Intonational-phrase-sensitive phones know whether they are final or nonfinal within an intonational phrase. Phrasal-prominence-sensitive phones know whether or not they have phrasal prominence. Word-level context is incorporated, for example, in audiovisual speech recognition models that represent pronunciation variability by way of within-word asynchrony among the targets achieved by the tongue, lips, and glottis/velum. Foot-sensitive phones represent the alternation among reduced, unreduced, and lexically stressed vowels. The syllable-sensitive phones described in this paper are, in fact, not phone models in the traditional sense at all; rather, they are better understood to be models of the consonant release and closure landmarks that initiate and terminate each syllable. Finally, two of the acoustic effects of disfluency have been represented: the unique acoustic characteristics of the phones in filled pauses, and the glottalization of phones in the final syllable of a reparandum. We report experimental results demonstrating that many of these context features may reduce the word error rate (WER) of a speech recognizer in at least one specified transcription task.[1] Computational complexity limitations, and training data limitations, have thus far prevented the integration of all proposed context features into any single ASR application.

[1] All of the experimental results described in this article have been previously published in technical reports or conference papers, but only the results of Section 2 have been previously published in professional journals; a more extensive description of one of the results of Section 6 is also currently under review. References to relevant technical reports and on-line documentation are provided in each section.

1 Introduction

This paper proposes using the prosodic hierarchy as an organizing framework for the sources of phonetic context information in both phone-based and articulatory-feature-based ASR. The goal of this introductory section is to adequately define the terms in the preceding sentence, and to give some of the reasons why we believe it to be a promising paradigm for ASR research.

An automatic speech recognizer is a search algorithm governed by a probability mass function (PMF). The PMF is an estimate of the probability, $P(W|X)$, that a talker has produced the word sequence $W = [w_1, \ldots, w_L]$ given that the acoustic signal has short-time spectra $X = [\vec{x}_1, \ldots, \vec{x}_T]$. The goal of the search algorithm is to find the W that maximizes $P(W|X)$:

$$\hat{W} = \arg\max_W P(W|X) \tag{1}$$

Researchers studying the "search problem" try to find an algorithm that maximizes $P(W|X)$ as fast as possible; researchers studying the "training problem" try to find a function $P(W|X)$ that is as accurate as possible. Because the field is specialized in this way, the accuracy of a speech recognizer is determined by the accuracy of its PMF model. The goal of accurate speech recognition is therefore equivalent to the goal of finding a function $P(W|X)$ such that, in all cases, the correct words (the words the talker actually said) are also the ones that maximize $P(W|X)$.

For computational reasons, Eq. 1 is usually rewritten as

$$\hat{W} = \arg\max_W \left(\frac{P(W)p(X|W)}{p(X)} \right) = \arg\max_W P(W)p(X|W) \tag{2}$$

The *language model* PMF $P(W)$ and the *acoustic model* probability density function (PDF) $p(X|W)$ are complicated functions with millions of trainable parameters. The acoustic model, $p(X|W)$, is parameterized by two fundamentally different types of parameters: *mode parameters* and *mixture parameters*. Mode parameters represent the mean and variance of an acoustic mode (a set of similar acoustic spectra that occur in similar linguistic contexts; a mode is usually modeled using a Gaussian distribution, therefore the mean and variance of the mode are sufficient statistics). Mixture parameters represent the different ways in which acoustic modes can be combined to form any given word sequence. There are two different types of mixture parameters: "mixture weights" specify the probability of disjunctive mode selection at a specified time, while "transition probabilities" specify the probability of any given mixture sequence.

Most words are infrequent, therefore it is impractical to learn the mode parameters and mixture parameters of every word directly from training data. Instead, most large-vocabulary speech recognizers simplify the mixture problem by defining a finite countable set of context-dependent, segmental units, intermediate between the word and the acoustic signal, called "phones." A well-designed phone set has the following properties:

- ACOUSTICALLY COMPACT: The phone label predicts the acoustic spectrum. In other words, given a phone label q_m at time t, the distribution $p(\vec{x}_t|q_m)$ of acoustic spectra has low entropy.
- PHONOLOGICALLY COMPACT: The word sequence predicts the phone sequence. In other words, given a word sequence $W = [w_1, \ldots, w_L]$, the distribution $P(Q|W)$ of possible phone sequences $Q = [q_1, \ldots, q_M]$ has low entropy.

It is not easy to define a set of phone labels that is both acoustically and phonologically compact. Orthographically identical phones may be acoustically disparate, e.g., there are acoustically important differences between the ten different productions of /t/ typical of the words "top," "tree," "stop," "steep," "felt," "bat," "bats," "batman," "butter," and "button" (Zue & Laferriere, 1979). Pronunciation depends on long-term context: an intonational-phrase-initial phone is different from an intonational-phrase-final phone, and a phone with phrasal prominence is different from a phone without phrasal prominence (Cole, Kim, Choi, & Hasegawa-Johnson, 2007). The relevant acoustic context is the entire utterance: prosodic phrases at the end of a prosodic group are shorter than prosodic-group-medial phrases (Tseng, Pin, Lee, Wang, & Chen, 2005).

In order to be acoustically compact, the phones used by an ASR must be context-dependent. The number of relevant contexts is quite large: it is not unusual for a typical ASR to use a phone inventory with tens of thousands of distinct phones. Each phone model represents a certain set of training examples: in order to specify the exact contexts in which those training examples occur, we need to define some notation. Standard notation makes a distinction between monophones, context-dependent phones (CD-phones), and states.

A "state" is an index into the table of parameterized acoustic probability distributions: given a unique state variable number or name, q, the recognizer is able to look up the parameters of the acoustic PDF $p(\vec{x}|q)$ in a parameter table. Most typically, the PDF $p(\vec{x}|q)$ is a diagonal covariance mixture Gaussian function (Juang, Levinson, & Sondhi, 1986). A mixture Gaussian PDF represents a D-dimensional acoustic feature vector, $\vec{x} = [x_1, \ldots, x_D]^T$, using a linear combination of K different Gaussian modes:

$$p(\vec{x}|q) = \sum_{k=1}^{K} c_{kq} \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{dkq}^2}} e^{-\frac{(x_d-\mu_{dkq})^2}{2\sigma_{dkq}^2}} \tag{3}$$

where the number of modes, K, and the dimension of the acoustic feature vector, D, are specified by the system designer, and all other parameters including the mixture weights c_{kq} and the mode parameters μ_{dkq} and σ_{dkq} are automatically learned from training data. Most ASR systems break each CD-phone into three temporally sequential states, so that the first state models the CD-phone onset and the third state models its offset (Jelinek, 1976).

Each CD-phone is a context-dependent variant of exactly one monophone. There are typically 48 monophones in English (Lee & Hon, 1989). The monophones correspond approximately, but not precisely, to phonemes. Non-phonemic monophones are created in order to represent unusually common and stable surface forms such as schwa (/AX/) and flap (/DX/). In this paper, monophones are expressed using two forms of notation: IPA notation (e.g., /noteʃən/) and two-letter ARPABET notation (e.g., `/N OW T EY SH AX N/`); to reduce confusion, the latter is written in capital Roman letters.[2] The contextual constraints acting on a CD-phone are specified using three delimiters: a preceding - denotes left context, a following + denotes right context, and a following _ denotes long-term context. For example, if the code `US` means "unstressed syllable," then the CD-phone `/AY-F+OW_US/` is a statistical model that has been trained to represent examples of the monophone /f/ occuring immediately after /ɑʲ/, immediately before /o/, in an unstressed syllable. States are specified by augmenting the CD-phone label with a number, e.g., `/AY-F+OW_US2/` is the second state of the CD-phone `/AY-F+OW_US/`, and $p(\vec{x}|\texttt{AY-F+OW\_US2})$ is the corresponding parameterized distribution of acoustic feature vectors.

As suggested by the notation, left context and right context are special, because they are used more universally than other types of context. A CD-phone dependent only on local left and right context (no long-term context) is called an n-phone (e.g., triphone, quinphone, or septphone; (Lee & Hon, 1989)). For example, the triphone `AY-F+OW` represents an /f/ produced at the center of the 3-phone sequence /ɑʲfo/, as in the word "triphone." If there are N monophones, then the number of possible n-phones is N^n. No reasonably-sized training corpus contains enough data to robustly train N^3 triphone models, therefore left-context phones and right-context phones with similar effects on the center phone are typically clustered together using a binary classification tree learned from training data (Odell, Woodland, & Young, 1994).

Standard phone notation, as introduced in the preceding paragraphs, suggests that the acoustic PDF $p(\vec{x}|q)$ is best indexed by some combination of the monophone label together with a series of context specifiers. Articulatory phonology (Browman & Goldstein, 1992) provides a quite different way of thinking about the effects of context on the acoustic correlates of a word. In articulatory phonology, a word is not stored in memory as a sequence of phones; instead, a word is stored as a partially sequenced set of intended articulatory gestures. The partial sequencing of gestures has been modeled as a graph of violable and sometimes conflicting alignment targets, mediated by a control algorithm (Nam & Saltzman, 2003); an alternative prior approach models phonology as a graph of pairwise temporal precedence relations between the onsets and/or offsets of articulatory states (Carson-Berndsen, 1999). Temporal overlap between competing gestures may block the perfect implementation of either gesture, leading to phonological assimilation or reduction. There is evidence to suggest that assimilation and reduction are planned rather than passive processes (e.g., Gomi and Kawato (1996) demonstrate that locality assimilation in manual reaching movements is centrally planned), therefore articulatory phonology posits a continuous-valued mental representation called the "tract

[2] For complete definitions of the ARPABET monophone inventory see, e.g., (Parsons, 1987; Lee & Hon, 1989; Young et al., 2002; Hasegawa-Johnson, 2005).

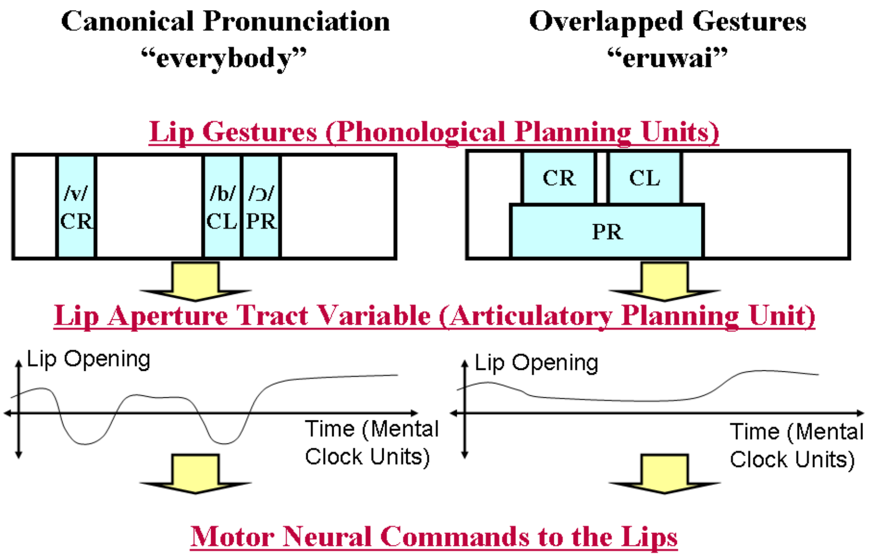

Canonical Pronunciation
"everybody"
Overlapped Gestures
"eruwai"
Lip Gestures (Phonological Planning Units)
/v/
CR
/b/
CL
/ɔ/
PR
CR
CL
PR
Lip Aperture Tract Variable (Articulatory Planning Unit)
Lip Opening
Time (Mental Clock Units)
Lip Opening
Time (Mental Clock Units)
Motor Neural Commands to the Lips

Figure 1: Overview of articulatory phonology. During rapid or casual speech, intended lip gestures are allowed to overlap in time (likewise tongue gestures, glottal gestures, et cetera). The phonological-to-articulatory transformation determines a target labial aperture ("tract variable") that serves as a compromise among the conflicting gestures. The articulatory-to-muscle-command transformation then generates motor unit commands that will achieve the target labial aperture. Lip gesture types shown in the figure are CLosed, CRitical (fricative), and PRotruded.

variable;" assimilation and reduction are planned during the mental transformation from gestures into tract variables (Saltzman & Munhall, 1989).

Fig. 1 provides a schematic of the way in which overlap among conflicting gestures may cause phonological reduction and assimilation: in this case, reduction of the word "everybody" to the casual form "eruwai" (ɛrʊwɑ^j) attested in a conversational telephone speech database (Livescu, 2005). The left half of Fig. 1 represents the phonological-to-articulatory planning process in the mind of a talker during production of the word "everybody" in citation form. Three intended lip gestures are shown for the word "everybody:" the CRitical gesture that defines the /v/, the CLosed gesture that defines the /b/, and the PRotruded gesture that defines the /ɔ/. Browman and Goldstein (1992) suggest that a gesture is specified in the lexical entry for any word if (and perhaps only if) it is necessary to distinguish that word from another word. Fig. 1 generalizes their claim slightly: we assume that the /ɔ/ inserts a LIP-PR gesture into this talker's lexical entry for the word "everybody," because /ɔ/ without the LIP-PR gesture would be /ʌ/, and despite the lack of any English word that would be confused with "everybody" if the LIP-PR gesture were omitted. During all other phones, Byrd and Saltzman (2003) suggest that the phonological-to-articulatory transform is driven by the influence of a "default gesture;" Fig. 1 assumes that the default gesture for the lips, during speech, is rather more open than closed. The "lip opening" tract variable smoothly interpolates between the target positions of the specified gestures; the articulatory-to-muscle-control transformation then determines muscle commands, to the several muscles of the lips, necessary to generate the desired labial aperture. Similar processes generate motor commands to the tongue, jaw, soft palate, larynx, and lungs.

The right half of Fig. 1 shows the speech planning process during rapid or casual speech; the lip gestures for the phones of "everybody" have been allowed to overlap.[3] Conflicting gestures specifying that the lips should be simultaneously CLosed and PRotruded can not be simultaneously satisfied, therefore the phonological-to-articulatory transformation works out a compromise: the lips will be narrow but open.

In a standard ASR, a "phone" is defined to be a monophone, modified by context specifications. In an ASR based on articulatory phonology, on the other hand, a "phone" may be defined as a vector of simultaneously

[3] For a comparable example of overlap among gestures within the same tract variable, see, e.g., Fig. 9 of (Byrd & Saltzman, 2003).

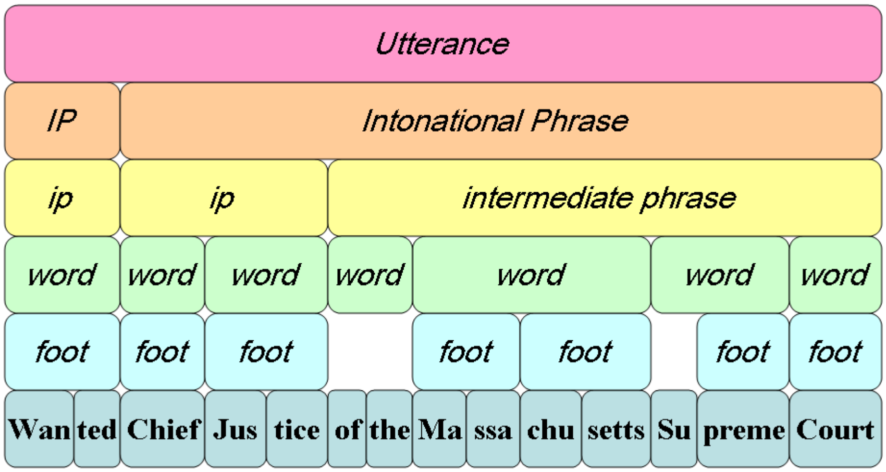


Figure 2: An example prosodic hierarchy with six levels, exemplified using a sentence from the Boston University Radio Speech Corpus (Ostendorf et al., 1995).

active gestures and tract variable settings. In articulatory phonology-based ASRs described in this paper, the vector of gestures and tract variables is never collapsed down to a single state variable. Instead, as proposed by Livescu and Glass (2004a), the ASR state variable, q, is replaced by a vector, $\vec{q} = [q_L, q_T, q_G]^T$ of quasi-independent articulatory features (AFs) representing the lips, tongue, and glottis/velum. Each articulatory feature is intended to be a summary of all gestures currently acting upon a particular named articulator. The vector of all currently active AFs serves as an index into a table of parameterized acoustic PDFs, $p(\vec{x}|\vec{q})$. This paper demonstrates in Secs. 3 and 5 that an AF vector is a useful replacement for the hidden state variable in ASR. An AF vector, however, has no explicit context specification: unlike the CD-phone label, the AF vector is not explicitly modified by triphone or prosodic context labels. It remains to be specified how we may represent prosodic and triphone context in an AF-based ASR.

Selkirk (Selkirk, 1981) proposes that any given phonological or phonetic sound pattern (that establishes a dependency between sounds or restricts the occurrence of a sound) must be defined in terms of relationships among the units at a specified level of the *prosodic hierarchy* (Fig. 2). Each level of the hierarchy is the relevant context for a particular set of phonetic and phonological sound patterns (processes and constraints). Some of the processes and constraints that have been proposed to operate at each level of the hierarchy include:

- Sound patterns bounded within the **Utterance** include the generation of turn-taking cues. It has been hypothesized that end-of-turn is cued by word choice, and also by modulation of some of the same acoustic features that are used to signal other types of prosodic juncture, e.g., pause, duration, and pitch (Local, Kelly, & Wells, 1986; Ferrer, Shriberg, & Stolcke, 2002; Gorman, Cole, Hasegawa-Johnson, & Fleck, 2007).

- Sound patterns bounded within the **Intonational Phrase** include boundary tones that mark the distinction between sentence types (e.g., question vs. statement), and possibly the specification of information structure distinctions such as theme vs. rheme (Steedman, 2000).

- Sound patterns bounded within the **Intermediate Phrase** include the assignment of phrasal stress and pitch accent, the downstepping of pitch accents in “list intonation” (Beckman & Elam, 1994; Yoon, 2007), the re-setting of pitch register, and the temporal regularization that defines rhythm (Kim, 2006).

- Sound patterns bounded within the **Prosodic Word** include the deletion or insertion of phones through processes related to syllabification, and many types of phonological feature assimilation.

- Sound patterns bounded within the **Foot** include the location of lexical stress, reduction and undershoot of both vowels and consonants, flapping, and possibly rhythmic adjustments in the direction of isochrony (Kim, 2006).

- Sound patterns bounded within the **Syllable** include the acoustic signaling of the phone itself. Stop consonants, for example, may be signaled by a consonant-vowel transition, a vowel-consonant transition, both, or neither.
- Disfluency may cause interruption and reset of any contiguous set of levels. For example, interruption of the word causes a word fragment; interruption of the intonational phrase causes pitch reset (Ostendorf, Shafran, Shattuck-Hufnagel, Carmichael, & Byrne, 2002; Cole et al., 2005).

This paper proposes using the prosodic hierarchy as an organizing framework for the sources of phonetic context information in both phone-based and articulatory-feature-based ASR. Specifically, this paper proposes two distinct methods for the explicit representation of prosodic context in ASR: one method that is appropriate for phone-based systems, and a quite different method that is appropriate for articulatory-feature-based systems.

Prosodic context may be incorporated into phone-based ASR by the use of long-term context specifications. Symbolically, the proposed scheme represents each phone as a vector of categorical features:

$$\text{phone} = \left[\begin{array}{c} \text{monophone label} \\ \text{syllable context features} \\ \text{foot context features} \\ \text{word context features} \\ \text{prosodic phrase context features} \\ \text{utterance context features} \\ \text{disfluency context features} \end{array}\right] \tag{4}$$

In implementation, the vector representation shown in Eq. 4 is collapsed into a single, rather long, CD-phone label. There are two ways to control the complexity of the resulting ASR. First, the differences among contextual variants of any given monophone may be constrained on the basis of phonetic knowledge, as described in Sec. 2. Second, the set of all CD variants of any given monophone may be clustered using the methods of (Odell et al., 1994). The methods for incorporating prosody into CD-phone-based ASR are reasonably well understood, and have been the subject of several published articles. Sec. 4 reviews the work of Bates and Ostendorf (2002, 2007), who use the methods of Eq. 4 to incorporate word-level, foot-level, and syllable-level context into the phone definition. Secs. 2 and 6 of this article describe our own previous work in this area, in which the methods of Eq. 4 are used to incorporate prosodic phrase context and disfluency context into the phone definition. In particular, Sec. 2 demonstrates that the use of a prosody-dependent acoustic model is not very effective unless combined with an explicit representation of prosody in the language model.

This paper proposes that prosodic context may be integrated into AF-based ASR by constraining the allowable sequences of articulatory features. The use of prosodic constraints in AF-based ASR is much less fully developed than the use of prosody-dependent phone models, but in general, it has the following properties. First, prosodic constituents (utterances, phrases, words, feet, or syllables) are specified symbolically in the language model and in the dictionary. Second, each prosodic constituent boundary imposes certain constraints on the articulatory feature trajectories as they cross the boundary. Sec. 3 describes, for example, a system in which all articulatory features must resynchronize at every word boundary, meaning that all articulators must simultaneously transition from one word to the next. We believe that this constraint is, in fact, too strict (examples of cross-word coarticulation are quite common in the literature (Beckman, 1989) and in our ASR training data (Greenberg, Hollenback, & Ellis, 1996)), but results in the phonology literature suggest that some less strict type of synchronization constraint is active at the word boundary and/or at the boundaries of intermediate or intonational phrases. Similarly, Sec. 5 describes a system in which a syllable is implicitly modeled as a sequence of phonetic landmarks: an optional consonant release, a required syllable nucleus, and an optional consonant closure. A transition of the articulatory features from a closed vocal tract state to an open vocal tract state, and back again, necessarily generates a sequence of three landmarks; the likelihood of that particular articulatory feature trajectory is then evaluated using a set of classifiers (support vector machines) trained to detect release, nucleus, and closure landmarks in the acoustic signal. Articulatory feature systems have not yet been designed to incorporate prosodic phrase context, utterance context, or disfluency context; Sec. 7 very briefly sketches methods that may be effective, in the future, for the incorporation of phrase-level prosodic context into AF-based ASR.

The phone-based and AF-based approaches described in this paper are intended to be complementary rather than contradictory. Hickok and Poeppel (2000, 2007) have recently argued, based on an extensive review of the neurophysiological literature, that the robustness of human speech perception is supported by the existence of at least three parallel neural pathways, any one of which is capable of independently accessing the mental lexicon. They demonstrate that the dorsal pathway is responsible for the transformation of acoustic percepts into signals that touch upon the articulatory motor pathway; they argue that signals in this path may then access the lexicon by way of articulation. The right ventral pathway, they argue, is capable of accessing the lexicon using only prosodic cues, e.g., syllable count and stress pattern, though the right ventral pathway can also make use of phone-level cues if available. Finally, the left ventral pathway accesses the mental lexicon with few steps, if any, intervening between sound patterns and stored word forms; it is to this pathway that Hickok and Poeppel attribute most of the classical results concerning phonological neighborhood effects on lexical access. Parallel computation is effective in ASR, too, and has been proven to be useful in a large number of recent papers (e.g., (Fiscus, 1997; Fosler-Lussier, 1999; Schwenk & Gauvain, 2000; Stolcke et al., 2001; Woodland, Hain, Evermann, & Povey, 2001; Martin & Przybocki, 2001)). Sec. 3 of this paper demonstrates a non-significant tendency for the lowest WER to be achieved by a system that combines the parallel outputs of a phone-based and an articulatory-feature based ASR.

2 Intonational and Intermediate Phrases

Intonational phrase (IP) boundaries are signalled by at least three types of cues: increased duration of phones in the rhyme of the phrase-final syllable (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992), a characteristic F0 movement called a boundary tone (Pierrehumbert, 1980), and increased glottalization of phrase-initial phones (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996). Within an intermediate phrase (ip), typically at least one word receives phrasal prominence: the stressed syllable of that word will be produced with greater vocal effort than prosodically unmarked phones of the same type, resulting in greater intensity and with increased duration that extends throughout the stress foot (Turk & Sawusch, 1997). There is often also a characteristic F0 movement associated with the accented syllable. This section considers, in particular, the use of increased phone duration as a cue for the detection of IP boundaries, and the use of F0 for the detection of phrasal prominence. Research in this area demonstrates that models of IP and ip context can reduce the word error rate (WER) of a speech recognizer.

By using speech data with manually transcribed intonational phrase boundaries and pitch accents, it is possible to train an automatic speech recognizer in which the prosodic context variable π_t for each phone takes one of four possible values: intonational phrase final vs. nonfinal, prominent vs. nonprominent (Chen et al., 2006). A phone in this system is defined to be phrase-final if it occurs in the rhyme of the syllable ending an intonational phrase, and nonfinal otherwise. A phone is defined to be prominent if it occurs in the lexically stressed syllable of a word marked as prominent in the prosodic phrase (i.e., marked with phrasal stress), and nonprominent otherwise. The prominent and nonprominent versions of each phone are allowed to differ only in the probability density function of an auxiliary normalized smoothed F0 observation, y_t; thus the joint probability density of the spectral envelope $\vec{x}_t$ and pitch y_t can be factored as $p(\vec{x}_t, y_t|c_t, \pi_t) = p(\vec{x}_t|c_t)p(y_t|c_t, \pi_t)$, where c_t is the triphone label. The spectral observation PDFs $p(\vec{x}_t, y_t|c_t, \pi_t)$ of the phrase-final and nonfinal versions of each triphone are not allowed to differ; only the model of phone duration is allowed to differ depending on intonational phrase position.

Table 1 shows WER of five different ASRs trained and tested using the Boston University Radio Speech Corpus (Ostendorf et al., 1995). The Radio Speech Corpus is a database of stories read, on the air and in the laboratory, by seven professional radio announcers. About 3.5 hours of speech have been prosodically transcribed using the ToBI (tones and break indices) prosodic transcription system (Silverman et al., 1992; Beckman & Hirschberg, 1994). A baseline ASR trained using 90% of the prosodicaly transcribed portion of the Radio Speech Corpus, and tested using the other 10%, achieved WER of 24.8%, shown in the first row of Table 1. By incorporating prosody-dependent acoustic models, WER was reduced to 24.0%.

The relationship among syntax, prosody, and the word string is modeled in our system by a prosody-dependent bigram language model. A prosody-dependent bigram is an estimate of $p(w_m, p_m|w_{m-1}, p_{m-1})$. The prosodic label p_m carries two types of information: the phrasal prominence/nonprominence of word w_m, and the position of w_m within an intonational phrase. There are eight possible settings of p_m: a word

Table 1: Word error rate (WER), prominence error rate (PER), and intonational phrase boundary error rate (BER, in percent) with five different combinations of acoustic model and language model. Chance performance is 44.6% PER, 15.6% BER.

Acoustic Model	Language Model	WER	PER	BER
Prosody Independent	Prosody Independent	24.8	44.6	15.6
Prosody Dependent	Prosody Independent	24.0	45.9	15.0
Prosody Independent	PD Bigram	24.3	23.1	14.5
Prosody Dependent	PD Bigram	23.4	20.3	14.3
Prosody Dependent	PD Semi-factored	21.7	20.3	14.2

may be prominent or nonprominent; the same word may be phrase-initial, phrase-final, phrase-medial, or it may be a one-word intonational phrase (both phrase-initial and phrase-final). The sequence $[p_{m-1}, p_m]$ takes on $|P|^2 = 64$ possible values, so in theory, a prosody-dependent bigram model learns 64 times as many parameters as a prosody-independent bigram model. In practice, most possible combinations of w_m and p_m never occur, so their probabilities are estimated by backing off to 1-gram and 0-gram (uniform) distributions; in our experiments, the actual parameter count of a prosody-dependent bigram model is slightly less than three times that of a prosody-independent bigram. A system using both prosody-dependent acoustic model and prosody-dependent language model, shown in the fourth row of Table 1, achieved WER of 23.4%—a significant reduction of word error rate in comparison to the baseline.

An empirically superior estimate of the prosody-dependent bigram probability may be trained by explicitly modeling the relationship between the prosodic tag, p_m, and the syntactic tag, s_m (Chen & Hasegawa-Johnson, 2003). The syntactic tagset used in our first-pass ASR specifies the part of speech of word w_m. By explicitly modeling syntactic tags, the prosody-dependent bigram probability may be written as

$$p(w_m, p_m | w_{m-1}, p_{m-1}) = \sum_{s_m, s_{m-1}} p(w_m, p_m, s_m, s_{m-1} | w_{m-1}, p_{m-1}) \quad (5)$$

$$\approx \sum_{s_m, s_{m-1}} p(p_m | s_m, s_{m-1}, p_{m-1}) p(s_m, s_{m-1} | w_m, w_{m-1}) p(w_m | w_{m-1}, p_{m-1}) \quad (6)$$

The approximation in Eq. 6 is valid if we assume that, first, prosody is independent of the word string given knowledge of syntax, and second, that the syntactic tags are independent of prosody given knowledge of the word string. The first term on the right-hand side of Eq. 6, $p(p_m | s_m, s_{m-1}, p_{m-1})$, may be robustly estimated from a relatively small corpus, because the syntactic tagset and the prosodic tagset are both much smaller than the vocabulary. The second term, $p(s_m, s_{m-1} | w_m, w_{m-1})$, is the probability that a word sequence (w_{m-1}, w_m) implements syntactic tag sequence (s_{m-1}, s_m); in our experiments we assumed this mapping to be deterministic. The third term in Eq. 6, $p(w_m | w_{m-1}, p_{m-1})$, is a prosody-dependent semi-bigram probability, and is estimated directly from the Radio Speech Corpus, using backed-off maximum likelihood estimation. A system using Eq. 6 to represent the language model achieved our lowest WER to date on the Boston University Radio Speech Corpus—21.7%.

3 The Word

Chomsky and Halle (1968) proposed that the domain of any given phonological process is bounded, with the domains of successive processes gradually expanding through a process of bracket erasure. In particular, they proposed that lexical stress assignment, phonotactics, and syllabification are determined within the boundaries of a lexical word. Selkirk (Selkirk, 1981) noted, however, that resyllabification often occurs across word boundaries, and proposed the "phonological word" to be the domain of syllabification. A phonological word is most often coterminous with a lexical word in English, but is quite frequently longer than a lexical word in Japanese (e.g., (Iwano & Hirose, 1999)) and Chinese (e.g., (Huang & Lee, 2006)), and is occasionally shorter than a lexical word in Spanish (Peperkamp, 1999). An example of a prosodic word composed of two lexical words is shown in Fig. 2, where the words "of the" have merged into a single prosodic word, allowing deletion of the final /v/ in "of," resulting in the open-syllable sequence /ə.ðə/. As in this example,

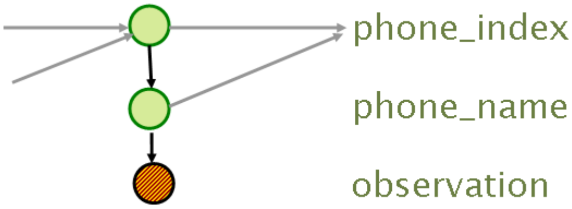


Figure 3: Dynamic Bayesian network (DBN) representation of a standard phone-based HMM speech recognizer.

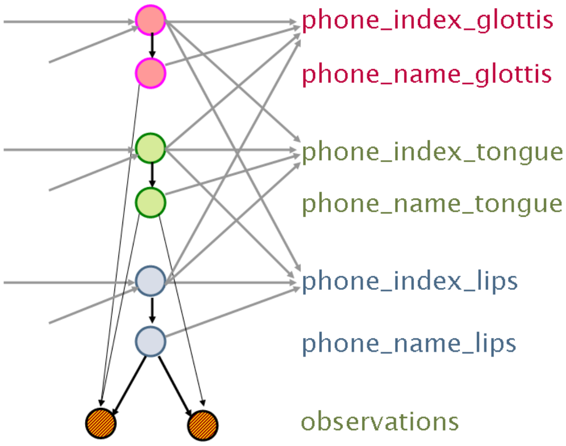


Figure 4: Dynamic Bayesian network (DBN) representation of a recognizer with three hidden state variables, separately representing the states of the lips, tongue, and glottis/velum.

resyllabification across a lexical word boundary may require phone deletion or substitution in order to avoid violating the phonotactic rules of the language. It is possible that, because of these resyllabification effects, phone deletion or substition effects across lexical word boundaries are more common within than between *prosodic* words, but we do not know of any published studies testing this hypothesis. Several published studies have proposed that cross-word coarticulatory effects are more common within than between prosodic phrases (Beckman, 1989; Beckman & Elam, 1994; Nakatani & Hirschberg, 1994).

Coarticulation and assimilation are modeled, in most modern ASR systems, by means of the n-phone abstraction (e.g., triphone or quinphone). An n-phone is a phoneme-length segment (a consonant or vowel), but the n-phone label depends on the phonological features of n consecutive segments: for example, the triphone AY-F+OW represents an /f/ produced at the center of the 3-phone sequence /ɑ^{j}fo/, as in the word "triphone" (Lee & Hon, 1989). In order to model the possibility that word boundaries may block coarticulation, many systems block the formation of triphones across a word boundary: for example, the /f/ in "my phone" may be represented by the biphone label F+OW instead of the triphone label AY-F+OW. Almost all modern systems use either cross-word triphones (in which triphone context extends across all word boundaries) or word-internal triphones (in which triphone context extends only within a lexical word), but Huang and Lee (2006) demonstrated that WER can be reduced by allowing cross-word triphones only when two lexical words are part of a single prosodic word.

Articulatory phonology has inspired a large number of recent ASR experiments (Richardson, Bilmes, & Diorio, 2000; Richmond, King, & Taylor, 2003; Livescu & Glass, 2004a). The model of Livescu and Glass (2004a), for example, factors the "phone" into a set of three to eight parallel "articulatory features" (AF), modeled as the hidden variables in a dynamic Bayesian network (DBN). For computational reasons, all published articulatory-phonology based ASR systems (including the system of (Livescu & Glass, 2004a), and the system described in this section) prohibit cross-word coarticulation. Asynchrony among the different articulators is allowed during a word. At a word boundary, however, every articulator is required to simulta-

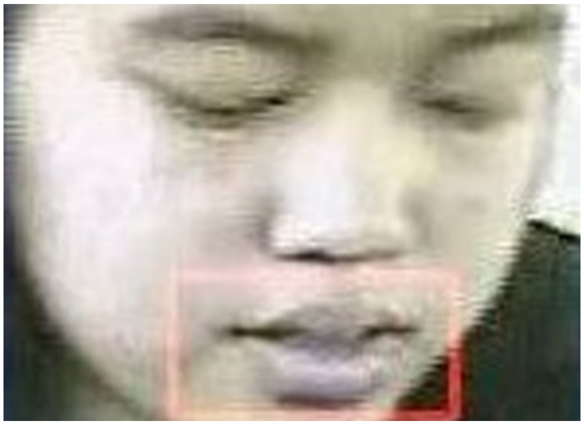


Figure 5: Asynchrony between audio and visual cues. The talker is preparing to begin saying the word "three;" there is not yet any audio signal. The tongue tip has been closed in preparation for the phoneme /θ/, and the lips have been rounded in preparation for the /r/.

neously change state. For example, at a hypothesized boundary between the words "two" and "three" with no intervening silence, the tongue closure and the glottal devoicing movement would be required to occur simultaneously; for computational reasons, the ASR would not be allowed to consider the hypothesis that tongue closure and glottal devoicing occur asynchronously. The prohibition of cross-word coarticulation in these systems has been implemented as a way of controlling computational complexity, and it is certainly too strict to represent real speech phenomena (examples of cross-word coarticulation are quite common in the literature (Beckman, 1989) and in our ASR training data (Greenberg et al., 1996)), but the work of Selkirk suggests that some kind of (looser) synchronization constraint may be appropriate at word boundaries, while the work of others (e.g., (Beckman, 1989)) suggests that prohibition of coarticulation across prosodic phrase boundaries would be appropriate.

The systems described in this section are based on the system of (Livescu & Glass, 2004a); all of these systems are implemented in GMTK (Zweig et al., 2002) using the notation of a dynamic Bayesian network or DBN. Fig. 3 shows a DBN representation of a standard hidden Markov model (HMM); Fig. 4 shows a DBN representation of a recognizer inspired by gestural phonology, with three different, conditionally independent articulators (the lips, tongue, and glottis/velum). The standard speech recognizer keeps track of two very different types of information about the phones at each time step: the `phone_index` specifies how far through the current word the talker has progressed, while the `phone_name` specifies which phone is actually being produced (which vowel or consonant it is). The `observation` (a perceptual LPC vector (Hermansky, 1990)) is dependent on the value of the `phone_name`. In the models proposed by Livescu and Glass (2004a), the `phone_name` is replaced by a set of three parallel labels: one label specifies the current state of the lips (wide, protruded, narrow, dental, closed, or silent), one label specifies the current state of the tongue (low back, high back, low front, high front, retroflex, palatal, palatal fricative, etc.), and the third label specifies the current state of the glottis and soft palate (unvoiced, voiced non-nasal, voiced nasal). The `observations` depend on the current settings of all three articulators.

It has long been recognized that the visual signal may convey evidence of inter-articulator asynchrony that is not obvious in the acoustic signal. Fig. 5, for example, shows a sample frame from the silence preceding the word "three:" although the acoustic signal is still silent, two of the three phones in the upcoming word are already visible in the talker's lips and tongue. It has been demonstrated many times that WER of an audiovisual speech recognizer may be reduced by explicitly modeling the asynchrony between audio and visual cues (e.g., (Chu & Huang, 2000; Neti et al., 2000; Zhang, 2000)). Asynchrony between audio and visual cues has most successfully been represented by the use of parallel HMMs: a "phoneme" model that generates audio feature observations, and a "viseme" model that generates video feature observations. One structure for managing the asynchrony between phoneme and viseme is the coupled HMM (CHMM) (Chu & Huang, 2000). As shown in Fig. 6, a CHMM is a DBN with two parallel sets of phone labels: a `phone_name_audio` representing the phone that is audible in the acoustic signal, and a `phone_name_video` representing the phone

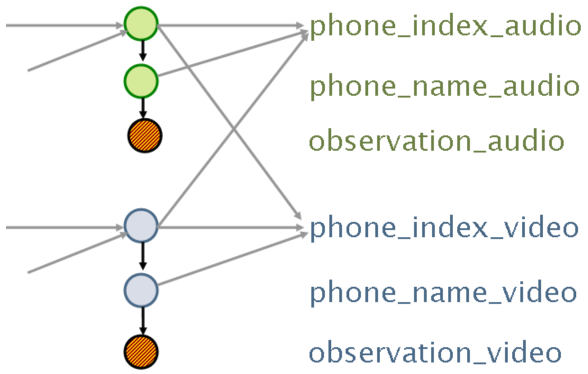


Figure 6: Coupled hidden Markov model (CHMM) designed to model asynchronies between the phone labels indicated by audio and visual speech observations. Each chain (audio and video) progresses through the same sequence of phones for any given word, but the two chains may progress at different rates.

Table 2: Word error rate, connected digit recognition, CUAVE development test data. Statistically significant differences are marked by a double line separating rows in the table.

System	WER
CHMM, up to 1 state of asynchrony allowed	22.8
Articulatory Feature system, 2 states asynchrony allowed	22.1
CHMM, up to 2 states of asynchrony allowed	21.8
ROVER system combination, three CHMM systems	20.1
ROVER system combination, two CHMM systems and one AF system	19.4

that is visible in the image sequence. These two phone labels may fall out of synchrony if, for example, the video images clearly show the tongue producing a /θ/, but the audio signal clearly contains only silence, as shown in Fig. 5.

In July 2006, we developed (Livescu et al., 2007) an audiovisual speech recognition system based on the gestural phonology model of Livescu and Glass. The system that we developed is shown in Fig. 4. That system was compared to the performance of the CHMM shown in Fig. 6, on the task of connected digit recognition from audiovisual recordings.[4] Training and test data were drawn from the CUAVE corpus (Patterson, Gurbuz, Tufecki, & Gowdy, 2002). Audio features included PLP coefficients, energy, deltas, and delta-deltas. Video features included the 35 lowest-order coefficients from a discrete cosine transform of the grayscale pixel values in a rectangle including the lips, and their deltas. Systems were trained using 60% of the available noise-free data. The number of Gaussians per mixture was increased until performance peaked on noise-free development test data (20% of the available data). Video and audio stream weights were then chosen in order to minimize WER on noisy development test data at six different SNRs (noise-free, 12dB, 10dB, 6dB, 4dB, and -4dB SNR), and the resulting WERs are reported in Table 2.

Results are shown in Table 2. The only statistically significant differences in this table are the difference between 20.1% WER and 21.8% WER, and the difference between 22.1% and 22.8% WER; all smaller differences are non-significant on this dataset. Trends shown in the table must be interpreted with caution, because they are not statistically significant, and because they have been obtained using development test data; confirmation of these results using independent evaluation test data was not completed. The trends shown in the table suggest, with the caveats already provided, that it would be worthwhile to pursue definitive support for the following conclusions. First, the CHMM seems to perform best when it is allowed to consider asynchrony between the states: as shown, allowing the audio and video phones to be asynchronous by two states (2/3 of a phone) is better than allowing only one state of asynchrony (1/3 of a phone). Similar results

[4] In standard ASR technical descriptions, "connected digits" are digits spoken with a silent pause after each word. Digits spoken with no pause between words are called "continuous." Connected speech recognition is generally considered to be easier than continuous speech recognition, but harder than isolated word recognition.

were achieved for the articulatory feature system. Second, it doesn't seem to matter very much exactly how the asynchrony is represented: the CHMM and the Articulatory-Feature system have almost identical word error rate (21.8% vs. 22.1%; the difference is not statistically significant, and reverses polarity in one of the noise conditions). Third, however, the two systems make slightly different types of errors, and therefore it is possible for the two systems to correct one another. If all three of these speech recognizers are allowed to vote in order to determine the output word string (using the ROVER paradigm (Fiscus, 1997)), word error rate is lower than the WER achieved by any one system alone. Furthermore, the ROVER combination of articulatory feature and CHMM systems has a tendency to be lower than the ROVER combination of three different CHMM systems (19.4% vs. 20.1% WER), suggesting that recognition accuracy may benefit from the use of two different methods to represent inter-articulator asynchrony.

All systems reported in Table 2 required the AF state variables to synchronize at every word boundary. It is common, in recent phone-based ASR systems, to allow two alternative pronunciations of each word: a version with cross-word triphones, and a version using only word-internal triphones (Woodland et al., 2001; Young et al., 2002). Similar experiments were attempted using the AF-based ASR: models were developed that allowed the articulatory features to be asynchronous across the boundary between a word and its neighboring silence. The model that allowed asynchrony across word boundaries was considerably more computationally complex than the models reported in Table 2. Because of the higher computational complexity, WER was only computed for the noise-free test condition; the resulting WER (7.5%) is significantly higher than the WER of any system in Table 2. Further research will seek to reduce the computational complexity and the WER of AF-based ASR with coarticulation across word boundaries.

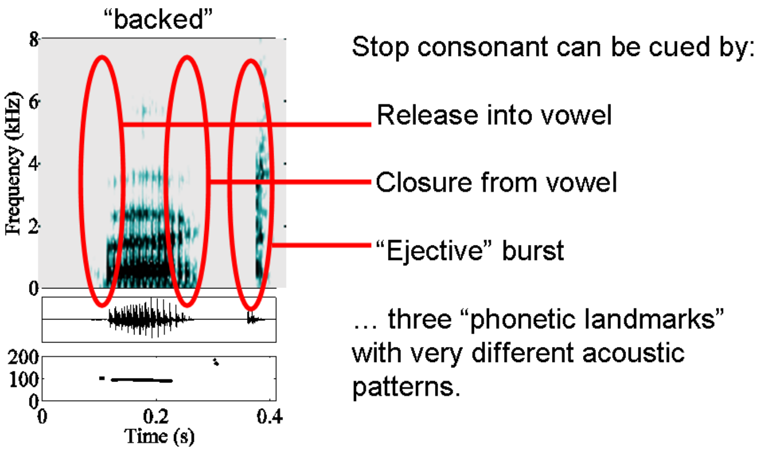
4 The Foot

The stress foot is the domain of lexical stress allocation, and of the strengthening or reduction of vowels and consonants (Turk & Sawusch, 1997; Kim, 2006). Lexical stress is deterministic, specified in the dictionary entry for all occurrences of a word, and therefore it is not difficult to use in ASR. In most standard English-language ASRs, for example, the dictionary entry for each word specifies whether any given vowel or alveolar stop should be implemented in reduced or unreduced form; reduced vowels are labeled as schwa (/AX/), and reduced intervocalic alveolar stops are labeled as flap (/DX/) (Lee & Hon, 1989). Some systems also distinctly model fronted schwa (/IX/) and/or nasal flap (/NX/) (Zue, Seneff, & Glass, 1990). These forms of reduction are hard-coded in the dictionary, and may be present in the dictionary regardless of whether or not the dictionary explicitly labels the location of lexical stress.

In the absence of phrasal prominence, it is not clear whether stress-related differences other than vowel reduction are useful for speech recognition. Van Kuijk and Boves (1999) found that unreduced lexically stressed and unstressed vowels, without pitch accent, did not differ significantly in pitch, energy, or duration, and hence were indistinguishable in an automatic speech recognition system. Bates and Ostendorf (2002, 2007), however, found that lexical stress can be used in automatic speech recognition as a form of optional context. In their study, triphone hidden Markov models were clustered into acoustically similar allophone clusters, as proposed by Odell, Woodland and Young (1994). In addition to the phoneme context questions proposed by Odell et al., however, Bates and Ostendorf also used questions about prosodic context (lexical stress, syllable position, and position in the word) to determine the clustering of allophones. The inclusion of prosodic context led to a statistically significant WER reduction. Hasegawa-Johnson (2006) has confirmed the results of Bates and Ostendorf using the training methods provided by a publicly available ASR toolkit.

5 The Syllable

Syllable context impacts the acoustic implementation of a phone more than context at any other level. Indeed, any given articulatory gesture may lead to radically different spectrotemporal patterns, depending on its syllable context. Consider, for example, the word "backed" (Fig. 7). This word contains three stop consonants; because of their relative positions in the syllable, the places of articulation of these three stops are communicated by three very different types of acoustic information. The place of the final /d/ is communicated by an ejective burst spectrum. The place of the /k/ is communicated by formant transitions during the last 70ms of the vowel. The place of the initial /b/ is communicated by both a turbulent burst and



Stop consonant can be cued by:

Release into vowel

Closure from vowel

"Ejective" burst

... three "phonetic landmarks" with very different acoustic patterns.

Figure 7: Redundancy of stop consonant landmarks: A stop consonant can be correctly recognized if a listener hears only the release (the /b/ in "backed"), only the closure (the /k/ in "backed"), or only an ejective release (the /d/ in "backed").

by formant transitions during the first 70ms of the vowel, but experiments with synthetic speech (Delattre, Liberman, & Cooper, 1955) and digitally modified natural speech (Nossair & Zahorian, 1991) have shown that either of these cues may be excised without impairing listeners' ability to understand the stop. The closure transition, burst spectrum, and release transition of a stop are thus redundant acoustic correlates; unambiguous presence of any one of these three acoustic patterns is enough to force listeners to hear the desired distinctive feature.

Context at the level of the syllable is modeled, explicitly or implicitly, in every modern ASR. Triphones, for example (Lee & Hon, 1989), implicitly distinguish between stop consonants that are signaled by the closure only (e.g., the /k/ in "backed," whose triphone representation is `/AE-K+D/`), the release only (e.g., the /k/ in "miscast," whose triphone representation is `/S-K+AE/`), or both (e.g., the /k/ in "backup," whose triphone representation is `/AE-K+AH/`). Since most triphones do not explicitly represent syllable boundary, however, some acoustically important effects are not coded by triphones, therefore it has been proposed that acoustic models should be sensitive to the locations of syllable boundaries (Greenberg, 1999). In the most extreme case, one may create an ASR that uses syllables or demi-syllables instead of phones as the fundamental building blocks of speech. The use of demisyllables as acoustic units is intuitively appealing, in part because it works so well in Chinese and Japanese. In English, however, the number of possible demisyllables is quite large, and the majority of possible demisyllables are rarely used, thus their acoustic correlates are not robustly represented in any reasonable-sized training corpus. Doddington et al. (1997) proposed solving the data sparsity problem by using syllabic acoustic units to augment a phone inventory rather than replacing it. Ganapathiraju et al. (2001) found that their best system included the following acoustic units: 200 monosyllable words, 632 common syllables, and triphones. In such a system, the "pronunciation" of any given word is given in terms of the largest available units: whole words if available, else syllables, else triphones.

Stevens et al. (Stevens, Manuel, Shattuck-Hufnagel, & Liu, 1992) proposed a different method for representing syllable context. In the "landmark-based speech recognizer" they proposed, phones are replaced by four different types of acoustic speech recognition units: consonant closure landmarks, consonant release landmarks, syllabic peak landmarks, and intervocalic glide landmarks. The set of English landmarks is reasonably small: depending on the way in which they are enumerated, one typically finds that there are fewer than 1000 acoustically distinct syllable-internal consonant-vowel and vowel-consonant biphones in English, and that all of them are well represented in a database the size of TIMIT (about 14 hours). To further simplify the task, Stevens et al. proposed detecting each landmark using a class-dependent modulation filtering algorithm, and labeling it using a series of binary distinctive feature classifiers. Landmark detection and distinctive feature classification algorithms have been developed using knowledge-based approaches (Espy-Wilson, 1994; Liu, 1995; Hasegawa-Johnson, 1996; Bitar & Espy-Wilson, 1996; Howitt, 2000; Chen, 2000;

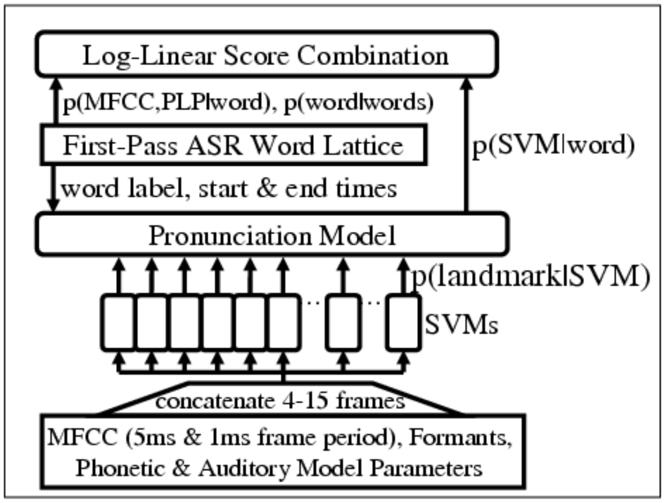

Log-Linear Score Combination
p(MFCC,PLP|word), p(word|words)
First-Pass ASR Word Lattice
p(SVM|word)
word label, start & end times
Pronunciation Model
p(landmark|SVM)
SVMs
concatenate 4-15 frames
MFCC (5ms & 1ms frame period), Formants,
Phonetic & Auditory Model Parameters

Figure 8: Schematic overview of landmark-based speech recognition systems implemented for large-vocabulary speech recognition by Hasegawa-Johnson et al. (2005)

Pruthi & Espy-Wilson, 2004), neural networks (Kirchhoff, Fink, & Sagerer, 2000; King & Taylor, 2000; Chang, Greenberg, & Wester, 2001), and support vector machines (SVMs) (Niyogi & Ramesh, 1998; Niyogi & Burges, 2002; Juneja & Espy-Wilson, 2003).

In July 2004, we trained and tested a number of different large-vocabulary continuous speech recognition (LVCSR) systems that fit the framework schematized in Fig. 8 (Hasegawa-Johnson et al., 2005). All LVCSR systems began with a high-dimensional multi-frame acoustic-to-distinctive feature transformation, implemented using SVMs trained to detect and classify landmarks. SVM inputs included MFCCs (computed using two different window lengths), formant frequencies and amplitudes (Zheng & Hasegawa-Johnson, 2004), knowledge-based acoustic parameters (Bitar & Espy-Wilson, 1996), and multiscale spectrotemporal rate features (STRFs) (Mesgarani, Slaney, & Shamma, 2004). Distinctive feature probabilities estimated by the support vector machines were then integrated using one of three different pronunciation models: a dynamic programming algorithm that assumes canonical pronunciation of each word, a DBN implementation of articulatory phonology, or a discriminative pronunciation model trained using the methods of maximum entropy classification. Log probability scores computed by these models were then combined, using log-linear combination, with the other word scores available in the lattice output of an HMM ASR, and the resulting combination scores were used to compute a second-pass speech recognition output.

A hybrid SVM-DBN landmark-based speech recognizer was created by combining the generative pronunciation model of (Livescu & Glass, 2004b) with the SVM acoustic observation probabilities described above. In the generative pronunciation model, hidden variables in a DBN represent features based on the tract variables of (Browman & Goldstein, 1992), including the locations and/or degrees of opening of the lips, tongue, and glottis/velum. Each word's baseform pronunciations are mapped to tract variable trajectories. The DBN allows the tract variables to go through their trajectories asynchronously (while enforcing some soft synchrony constraints, encoded as distributions over degrees of asynchrony). The system developed in this way is similar to that shown in Fig. 4, with two key differences. First, the lips, tongue, and glottis/velum are allowed to take on "surface" values that differ from their canonical or "underlying" phone targets: for example, the variable phone_name_lips is divided into two hidden variables called, respectively, phone_name_lips_underlying and phone_name_lips_surface. Second, instead of PLP observations, the landmark-based speech recognizer observes the classification posterior probabilities computed by forty different SVMs trained to detect and classify landmarks.

Table 3 shows a sample of the word error rates obtained with this system on a three-speaker subset of the RT03 development test set. The baseline system in these experiments was the SRI EARS large vocabulary speech recognizer as of 2003 (Stolcke et al., 2003). It is worth noting that the WER of any speech recognizer is a moving target: the WER of the 2005 SRI system was approximately half that of the 2003 system. All

Table 3: Word error rates (%) in lattice rescoring experiments on a three-speaker subset of the RT03 development set. The last line of the table shows the WER achieved when the DBN observes only those SVMs whose per-frame binary classification accuracy exceeds a reasonable threshold.

System setup	WER
Baseline	27.7
SVM-DBN, all SVMs	27.3
SVM-DBN, high-accuracy SVMs only	27.2

rescoring experiments combined the log likelihoods of the SRI recognizer with log likelihoods of the DBN. Two rescoring experiments are reported in the table. In the first experiment, the DBN observes outputs of all SVM-based landmark detectors and classifiers. In the second experiment, the DBN observes the outputs of only the SVMs whose per-frame classification accuracy exceeds some reasonable threshold. The proposed methods show a trend (not statistically significant) toward reduction of WER on this development test dataset. Confirmation of these results using independent evaluation test data was not completed.

6 Disfluency

Disfluency can change the acoustic implementation of a phone, therefore the minimization of WER requires some representation of disfluency. Fortunately, disfluency is relatively easy to identify, in the following senses. First, linguistically naive transcribers are able to locate filled pauses and the interruption point of a repair or repetition disfluency with high levels of inter-transcriber agreement (Shriberg, 2000; Meteer & Taylor, 1995). Second, most disfluencies follow relatively stylized patterns of repair, repetition, and filled pause, and most disfluencies are therefore relatively easy to detect from an orthographic transcription of speech (Baron, Shriberg, & Stolcke, 2002; Gupta, Bangalore, & Rahim, 2002; Kim, Schwarm, & Ostendorf, 2004; Lendvai, van den Bosch, & Krahmer, 2003). The key difficulties in the transcription of disfluency are: (1) if disfluency is not adequately modeled by the phone set of an ASR, disfluencies will be mis-transcribed as if they were fluent speech, causing a large number of speech recognition errors (Adda-Decker et al., 2003; Aylett, 2003; Rose & Riccardi, 1999), (2) all of the previous discussion refers to the most common patterns of disfluency, but some types of disfluency do not follow these patterns and are therefore difficult to transcribe (Shriberg, 2001).

Fig. 9 shows a disfluency with a double reparandum: "I, I, one of the things I..." Fig. 9, like the remainder of this section, adopts the disfluency annotation system of Heeman and Allen (Heeman & Allen, 1999). In their annotation system, the words being corrected are called the "reparandum" or REP, the correction is called the "alteration" (ALT), and filled pauses or meta-dialog are called the "edit" (EDT). In Fig. 9, the first reparandum is repeated, then finally repaired by the alteration. As shown, we find that most repair and repetition disfluencies in Switchboard contain no verbal EDT segment—many REP segments end in glottalization and/or elongation, but rarely in a verbal EDT segment. Conversely, most verbal EDT segments take the form of explicit filled pauses, most typically "uh" or "um" (Clark & Fox Tree, 2002).

Disfluency is common in conversational speech. Of 1100 words we have transcribed (Yoon, Chavarria, Cole, & Hasegawa-Johnson, 2004; Cole et al., 2005), 40 are part of a reparandum, 37 are filled pauses, and 41 are part of an alteration, thus 10% of the words we have transcribed are part of a disfluency. This estimate is higher than most published estimates, perhaps because we include all words that are part of the reparandum or alteration, but most published studies estimate that at least 5% of the words in Switchboard are part of a disfluency (e.g., (Shriberg, 2001)).

REP and ALT segments are not transcribed in most speech recognition training corpora, therefore it is difficult to train an ASR model of all aspects of disfluency. Two aspects of disfluency, however, are commonly transcribed in all speech recognition training corpora. First, filled pauses are usually labeled with unique lexical tokens: in the Switchboard corpus, for example (Godfrey, Holliman, & McDaniel, 1992), the words "UH" and "UM" are uniquely used to label filled pauses. Second, word fragments are often uniquely labeled. In Switchboard, for example, annotations specify the word that the talker was apparently trying to say (in the judgment of the transcriber); the unsaid portion is enclosed in brackets, e.g., the phone sequence /juni?/

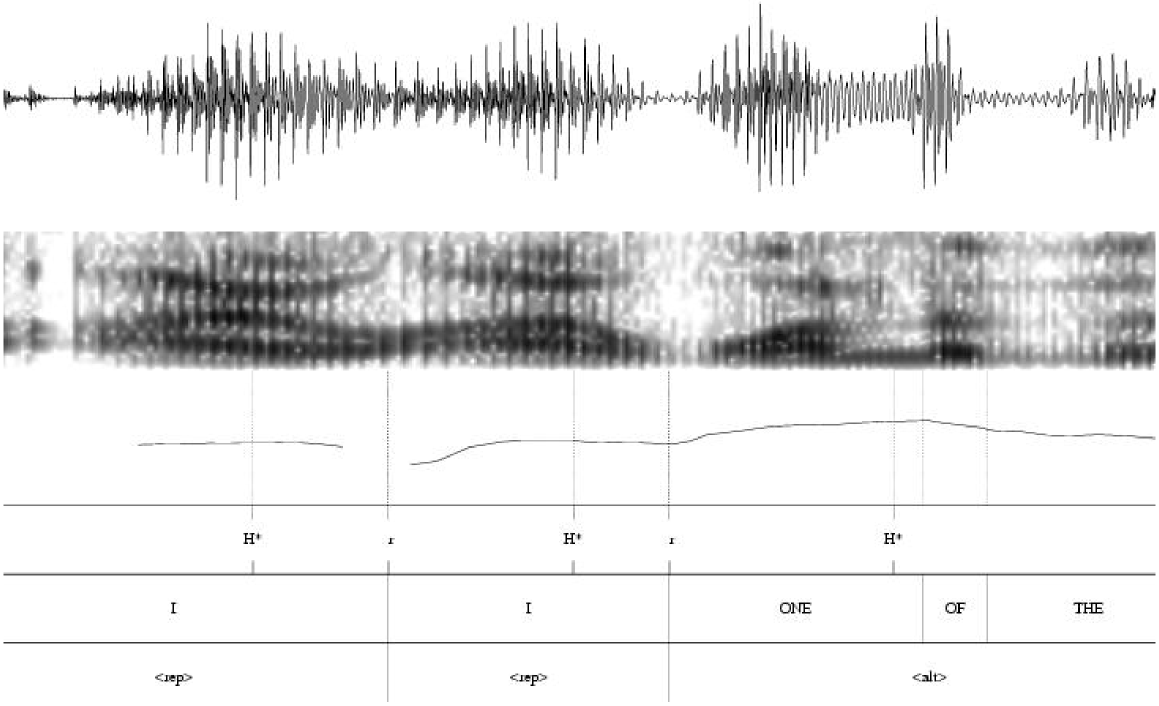


Figure 9: Transcription of prosody and disfluencies in the phrase "I, I, one of the..."

might be transcribed using the word fragment "UNI[QUE]." Word fragments occur almost exclusively at the end of a disfluency reparandum, therefore word fragment labels specify the end (but not the beginning) of some (but not all) disfluency reparanda.

Filled pauses may be treated as regular lexical tokens in an ASR language model, forcing the language model to separately learn the lists of words which typically precede an "UH" or an "UM." Unlike the language model, the acoustic model of an ASR may benefit by giving "UH" and "UM" special treatment. The vowel in "UH" is acoustically similar to the vowel /ʌ/ in content words like "TUG," but the /ʌ/ of "UH" is usually weaker and longer. Similarly, the word "UM" is often produced with a drawn-out, low-intensity /m/. If the word "UM" is modeled by the phone sequence /ʌm/, then the aberrant statistics of the /m/ in "UM" will reduce the precision of the statistical model of /m/: because the phone model of /m/ is being used to represent both fluent and disfluent productions, it fails to compactly represent either. For these reasons, Greenberg, Hollenback, and Ellis (1996) proposed representing the words "UH" and "UM" with the unique filled-pause phones /PV/ ("pause vowel") and /PN/ ("pause nasal").

The end of a REP segment—especially a REP that ends in a word fragment—is often glottalized. In Fig. 9, for example, glottalization is visible at both interruption points: the first REP segment ends in low-pitched creaky voicing, while the second REP segment ends in a glottal stop. Yoon et al. (Yoon, Zhuang, Cole, & Hasegawa-Johnson, 2006) have shown that WER of an ASR may be reduced by using an automatically labeled "creaky" vs. "modal" distinction as part of the definition of a phone.

7 Conclusions and Future Work

This paper has proposed using the prosodic hierarchy as an organizing framework for the sources of acoustically salient context information in ASR. Specifically, we have discussed five experimental systems, each of which divides the phone inventory into two or more subcategories as specified by the following prosodic and disfluency context features:

1. Position within intonational phrase (final vs. nonfinal)

2. Phrasal prominence (prominent vs. nonprominent)

3. Position within prosodic word (initial, medial, or final)

4. N-phone context (manner, place, and voicing of the preceding and following phones)

5. Lexical stress (primary stress, reduced, neither)

6. Syllable position (consonant release, consonant closure, syllabic nucleus, or intervocalic glide)

7. Fluency (filled pause vs. non-pause)

8. Voicing (creaky vs. modal)

Equation 4 suggests that all of the features above should be used to define a phone inventory. It is impractical, however, to divide a small speech training corpus into mutually exclusive subsets representing every possible combination of the features listed above. Instead, it is necessary to find some method of computing, and applying, an estimate of the specific acoustic transformations that relate one prosodic context to another.

Sec. 2 proposed using phonetic knowledge to define the most important acoustic differences among prosodic contexts. For example, in that section, the models of phrasally prominent and non-prominent examples of the same underlying phoneme are tied together in all acoustic dimensions but F0. Similarly, phrase final and nonfinal phones are tied together in all acoustic dimensions but duration.

Sec. 3 reviewed a common "tree-based splitting" approach to triphone context features, first proposed by Odell, Woodland, and Young (1994). In that standard approach, the phone inventory of an ASR system is created through a tree-structured series of binary divisions of the training data. Each binary split is selected, from a list of candidate binary context features, in order to make the leaves of the new tree as acoustically compact as possible. The splitting process continues while each leaf of the tree contains a sufficient number of training examples. Bates and Ostendorf (2002, 2007) proposed using a similar binary splitting method to model the acoustic salience of arbitrary prosodic context variables including syllable position, word position, and lexical stress. Borys (2003) proposed using the same method to model the acoustic salience of intonational phrase position and phrasal prominence. Yoon et al. (2006) proposed using the same method to model the acoustic salience of voice quality labels.

Exhaustive splitting and tree-based splitting methods both work from the assumption that the "context-dependent phone" is an indivisible unit. Livescu and Glass (2004b) have suggested, rather, that the scalar phone label should be split into a vector of AF labels, each representing the targets achieved by one articulator. Browman and Goldstein (1992) go one step farther, arguing that the phone should be replaced by three distinct set representations at each time t: a set of "gestures" that are intended or desirable at time t, a vector of "tract variables" that have been planned for production at time t, and a vector of articulator positions that are actually produced at time t. Most implemented computational models of articulatory phonology posit that the mapping from tract variables to articulator positions is usually trouble-free (in speech without pathology): most pronunciation variability comes from the mapping between gestures and tract variables.

All of the context variables discussed in this paper can be re-written in terms of articulatory phonology. For example, articulatory phonology greatly simplifies the representation of triphone context: All of the effects of triphone context are represented, in articulatory phonology, by the temporal overlap of competing gestures. The blocking of coarticulation across word or phrase boundaries may be represented, as suggested in Sec. 3, by forcing the gestures or tract variables to re-synchronize at the boundaries of prosodic words or phrases. The effects of syllable context may be represented, as suggested in Sec. 5, by developing distinct SVM or neural network classifiers designed to detect and classify the release and closure landmarks associated with any particular articulator.

Future work will try to develop comparable representations, in terms of articulatory phonology, for the effects of prosodic phrase context, prosodic group context, and disfluency. A promising method is suggested by the work of Byrd and Saltzman (2003). Byrd and Saltzman developed, based on the work of (Saltzman & Munhall, 1989), an algorithm for synthesizing articulator kinematics from hypothesized articulatory gestures. In their model, phrase boundaries are modeled by a π_T gesture (a "lengthening" gesture (Beckman & Edwards, 1990)), whose function is to slow down the clock controlling the mapping between gestures and tract variables. Similarly, prominence is modeled by a π_S gesture (a "strengthening" gesture (Fougeron

& Keating, 1997)), whose function is to increase the magnitude of all tract variable excursions during its period of activity. There is a natural mapping between the context variables considered in this paper and the π_S and π_T gestures of Byrd and Saltzman: lexical stress and phrasal prominence are different types of π_S gesture, while utterance, intonational phrase, and intermediate phrase boundaries each generate a different type of π_T gesture. The Articulatory Feature (AF) models of Secs. 3 and 5 provide a good starting point for the implementation of a prosody-dependent articulatory feature ASR, e.g., it may be possible to simply add two more hidden state variables representing π_S and π_T. In order for these ideas to become useful in automatic speech recognition, the biggest remaining unsolved problem seems to be the creation of a probabilistic representation of the "lengthening" and "strengthening" functions—that is to say, we need to somehow represent "lengthening" and "strengthening" as learnable context-dependent transformations of the mode parameters or mixture parameters of a statistical ASR.

8 Acknowledgements

This material is based upon work supported by the the National Science Foundation (NSF) under grants 04-14117 (Hasegawa-Johnson, Cole, and Shih) and 01-21285 (Jelinek and Khudanpur). Any opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Adda-Decker, M., Habert, B., Barras, C., Adda, G., Mareuil, P. B. de, & Paroubek, P. (2003). A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. In *ISCA workshop on disfluency in spontaneous speech.* Göteberg, Sweden.

Aylett, M. P. (2003). Disfluency and speech recognition profile factors. In *ISCA workshop on disfluency in spontaneous speech.* Göteberg, Sweden.

Baron, D., Shriberg, E., & Stolcke, A. (2002). Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *Proc. internat. conf. spoken language processing.* Orlando.

Bates, R., & Ostendorf, M. (2002). Modeling pronunciation variation in conversational speech using prosody. In *ISCA workshop on pronunciation modeling and lexical access.* Baltimore.

Bates, R. A., Ostendorf, M., & Wright, R. A. (2007). Symbolic phonetic features for modeling of pronunciation variation. *Speech Communication*, *49*, 83-97.

Beckman, M. E. (1989). Timing models for prosody and cross-word coarticulation in connected speech. In *Proc. of the darpa speech recognition workshop* (p. 12-21).

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. Beckman (Eds.), *Between the grammar and physics of speech: Papers in laboratory phonology I* (p. 152-178). Cambridge: Cambridge University Press.

Beckman, M. E., & Elam, G. A. (1994). *Guidelines for ToBI labelling* (Tech. Rep.). Ohio State University. (http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html)

Beckman, M. E., & Hirschberg, J. (1994). *The ToBI annotation conventions* (Tech. Rep.). Ohio State University and Columbia University.

Bitar, N., & Espy-Wilson, C. (1996). A knowledge-based signal representation for speech recognition. In *Proc. ICASSP* (p. 29-32). Atlanta.

Borys, S. (2003). The importance of prosodic factors in phoneme modeling with applications to speech recognition. In *HLT/NAACL student session.* Edmonton.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155-180.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *J. Phonetics*, *31*, 149-180.

Carson-Berndsen, J. (1999). *Time map phonology: Finite state models and event logics in speech recognition.* Kluwer Academic Publishers.

Chang, S., Greenberg, S., & Wester, M. (2001). An elitist approach to articulatory-acoustic feature classification. In *Proc. EUROSPEECH.* Aalborg, Denmark.

Chen, K., & Hasegawa-Johnson, M. (2003). Improving the robustness of prosody dependent language modeling based on prosody syntax cross-correlation. In *IEEE workshop on automatic speech recognition and understanding (ASRU).* Virgin Islands.

Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S.-S., Cole, J., & Choi, J.-Y. (2006). Prosody dependent speech recognition on radio news. *IEEE Trans. Speech and Audio Processing*, *14*(1), 232-245.

Chen, M. (2000). Nasal landmark detection. In *Proc. internat. conf. spoken language processing* (p. 636-639). Beijing.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York, NY: Harper and Row.

Chu, S., & Huang, T. S. (2000). Bimodal speech recognition using coupled hidden Markov models. In *Proc. internat. conf. spoken language processing.* Beijing.

Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*, 73-111.

Cole, J., Hasegawa-Johnson, M., Shih, C., Kim, H., Lee, E.-K., Lu, H. yi, Mo, Y., & Yoon, T.-J. (2005). Prosodic parallelism as a cue to repetition and error correction repair disfluency. In *ISCA workshop on disfluency in spontaneous speech.* Aix-en-Provence.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from radio news speech. *J. Phonetics*, *35*, 180-209.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*(4), 769-773.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *J. Phonetics*, *24*, 423-444.

Doddington, G., Corrada, A., Ganapathiraju, A., Goel, V., Wheatley, B., Kirchhoff, K., Ordowski, M., & Picone, J. (1997). *Syllable-based speech processing* (Tech. Rep. No. WS97). Johns Hopkins Center for Language and Speech Processing.

Espy-Wilson, C. (1994). A feature-based semi-vowel recognition system. *Journal of the Acoustical Society of America*, *96*(1), 65-72.

Ferrer, L., Shriberg, E., & Stolcke, A. (2002). Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In *Proc. internat. conf. spoken language processing.* Denver.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE workshop on automatic speech recognition and understanding (ASRU).* Santa Barbara.

Fosler-Lussier, E. (1999). Contextual word and syllable pronunciation models. In *IEEE workshop on automatic speech recognition and understanding (ASRU).*

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*(6), 3728-3740.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., & Doddington, G. R. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, *9*(4), 358-366.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *Proc. ICASSP* (p. 517-520).

Gomi, H., & Kawato, M. (1996). Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement. *Science*, *272*, 117-120.

Gorman, K., Cole, J., Hasegawa-Johnson, M., & Fleck, M. (2007). *Automatic detection of turn-taking cues in spontaneous speech using prosodic features.* Paper presented at the 81st annual meeting of the linguistic society of america, Anaheim, CA.

Greenberg, S. (1999). Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*, 159-176.

Greenberg, S., Hollenback, J., & Ellis, D. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proc. internat. conf. spoken language processing.* Philadelphia.

Gupta, N. K., Bangalore, S., & Rahim, M. (2002). Extracting clauses for spoken language understanding in conversational systems. In *Proc. internat. conf. spoken language processing.* Orlando.

Hasegawa-Johnson, M. (1996). *Formant and burst spectral measurements with quantitative error models for speech sound classification.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Hasegawa-Johnson, M. (2005, May). *Speech tools minicourse.* (Available on the web at http://www.isle.uiuc.edu/courses/minicourse/index.html)

Hasegawa-Johnson, M. (2006, Jan). *HTK study group: Recognizer training and testing methods.* (Available on the web at http://www.isle.uiuc.edu/courses/htk/index.html)

Hasegawa-Johnson, M., Baker, J., Greenberg, S., Kirchhoff, K., Muller, J., Sönmez, K., Borys, S., Chen, K., Juneja, A., Livescu, K., Mohan, S., Coogan, E., & Wang, T. (2005). *Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop* (Tech. Rep. No. WS04). Johns Hopkins University Center for Language and Speech Processing. (http://www.clsp.jhu.edu/ws2004/groups/ws04ldmk/ws04ldmk_final.pdf)

Heeman, P. A., & Allen, J. F. (1999). Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, *25*(4).

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, *87*(4), 1738-1752.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 1463-7.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews*, *8*, 393-402.

Howitt, A. W. (2000). Vowel landmark detection. In *Proc. internat. conf. spoken language processing.* Beijing.

Huang, J.-T., & Lee, L.-S. (2006). Detection of prosodic words in Mandarin Chinese. In *Proc. speechprosody.* Dresden.

Iwano, K., & Hirose, K. (1999). Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition. In *Proc. ICASSP.* Phoenix.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proc. IEEE*, *64*(4), 532-556.

Juang, B. H., Levinson, S. E., & Sondhi, M. M. (1986). Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Trans. on Information Theory*, *32*(2), 307-309.

Juneja, A., & Espy-Wilson, C. (2003). Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. In *Proc. internat. joint conf. neural networks (ijcnn).* Portland, OR.

Kim, H. (2006). *Rhythmic shortening in American English: Effect of prosodic phrase structure.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Kim, J., Schwarm, S. E., & Ostendorf, M. (2004). Detecting structural metadata with decision trees and transformation-based learning. In *Proc. hlt/naacl.* Boston.

King, S., & Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language*, *14*(4), 333-354.

Kirchhoff, K., Fink, G., & Sagerer, G. (2000). Conversational speech recognition using acoustic and articulatory input. In *Proc. ICASSP.* Istanbul, Turkey.

Lee, K.-F., & Hon, H.-W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoustics, Speech, and Signal Processing*, *37*(11), 1641-1648.

Lendvai, P., van den Bosch, A., & Krahmer, E. (2003). Memory-based disfluency chunking. In *ISCA workshop on disfluency in spontaneous speech.* Göteborg, Sweden.

Liu, S. A. (1995). *Landmark detection for distinctive feature-based speech recognition.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Livescu, K. (2005). *Feature-based pronunciation modeling for automatic speech recognition.* Unpublished doctoral dissertation, MIT.

Livescu, K., Çetin, Özgür., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, A., Lal, P., Yung, L., Bezman, A., Dawson-Hagerty, S., Woods, B., Frankel, J., Magimai-Doss, M., & Saenko, K. (2007). *Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 JHU summer workshop final report* (Tech. Rep. No. WS06). Johns Hopkins University Center for Language and Speech Processing.

Livescu, K., & Glass, J. (2004a). Feature-based pronunciation modeling for speech recognition. In *Proc. hlt/naacl.* Boston.

Livescu, K., & Glass, J. (2004b). Feature-based pronunciation modeling with trainable asynchrony probabilities. In *Proc. interspeech.* Jeju Island, Korea.

Local, J., Kelly, J., & Wells, W. (1986). Towards a phonology of conversation: Turn-taking in tyneside english. *J. Phonetics*, *22.*

Martin, A., & Przybocki, M. (2001). The 2001 nist evaluation for recognition of conversational speech over the telephone. In *Nist 2001 workshop on speech transcription.*

Mesgarani, N., Slaney, M., & Shamma, S. A. (2004). Speech discrimination based on multiscale spectrotemporal features. In *Proc. ICASSP.* Montreal.

Meteer, M., & Taylor, A. (1995). *Dysfluency annotation stylebook for the switchboard corpus* (Tech. Rep.). Linguistic Data Consortium.

Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, *95*(3), 1603-1616.

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In *International conference on the phonetic sciences.* Barcelona.

Neti, C., Luettin, G. P. J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., & Zhou, J. (2000). *Audio-visual speech recognition: Final report* (Tech. Rep. No. WS00). Johns Hopkins University Center for Language and Speech Processing.

Niyogi, P., & Burges, C. (2002). *Detecting and interpreting acoustic features by support vector machines* (Tech. Rep. No. 2002-02). University of Chicago Computer Science Dept.

Niyogi, P., & Ramesh, P. (1998). Incorporating voice onset time to improve letter recognition accuracies. In *Proc. ICASSP* (p. 13-16). Seattle.

Nossair, Z. B., & Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, *89*(6), 2978-2991.

Odell, J. J., Woodland, P. C., & Young, S. J. (1994). Tree-based state clustering for large vocabulary speech recognition. In *Proc. internat. sympos. speech, image process. and neural networks* (p. 690-693). Hong Kong.

Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1995). *The Boston university radio speech corpus.* Linguistic Data Consortium.

Ostendorf, M., Shafran, I., Shattuck-Hufnagel, S., Carmichael, L., & Byrne, W. (2002). A prosodically labeled database of spontaneous speech. In *Proc. isca tutorial and research workshop on prosody in speech recognition and understanding.* Red Bank, NJ.

Parsons, T. (1987). *Voice and speech processing.* New York: McGraw-Hill.

Patterson, E., Gurbuz, S., Tufecki, Z., & Gowdy, J. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Proc. ICASSP.* Orlando.

Peperkamp, S. (1999). Prosodic words. *Glot International*, *4*(4), 15-16.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Pruthi, T., & Espy-Wilson, C. Y. (2004). Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, *43*(3), 225-240.

Richardson, M., Bilmes, J., & Diorio, C. (2000). Hidden-articulator Markov models: performance improvements and robustness to noise. In *Proc. internat. conf. spoken language processing.* Beijing.

Richmond, K., King, S., & Taylor, P. (2003). Modeling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, *17*(2-3), 153-172.

Rose, R. C., & Riccardi, G. (1999). Modeling disfluency and background events in ASR for a natural language understanding task. In *Proc. ICASSP.* Phoenix.

Saltzman, E. L., & Munhall, K. J. (1989). A dynamical approach to gestural patterning in speech production. *Haskins Laboratories Status Report on Speech Research*, *SR-99/100*, 38-68.

Schwenk, H., & Gauvain, J.-L. (2000). Combining multiple speech recognizers using voting and language model information. In *Proc. internat. conf. spoken language processing* (p. 915-918). Beijing.

Selkirk, E. O. (1981). *The phrase phonology of English and French.* Bloomington, Indiana: Indiana University Linguistics Club.

Shriberg, E. (2000). Disfluencies in Switchboard. In *Proc. internat. conf. spoken language processing.* Beijing.

Shriberg, E. (2001). To 'errr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153-164.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proc. internat. conf. spoken language processing.* Banff.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, *31*(4), 649-689.

Stevens, K. N., Manuel, S. Y., Shattuck-Hufnagel, S., & Liu, S. (1992). Implementation of a model for lexical access based on features. In *Proc. internat. conf. spoken language processing* (Vol. 1, p. 499-502). Banff, Alberta.

Stolcke, A., Abrash, V., Franco, H., Gadde, R. R., Shriberg, E., Sonmez, K., Venkataraman, A., Vergyri, D., & Zheng, J. (2001). The sri march 2001 hub-5 conversational speech transcription system. In *Nist workshop on speech transcription.*

Stolcke, A., Franco, H., Gadde, R., Graciarena, M., Precoda, K., Venkataraman, M., Vergyri, D., Wang, W., Zheng, J., Huang, Y., Peskin, B., Bulyko, I., Ostendorf, M., & Kirchhoff, K. (2003). Speech-to-text research at sri-icsi-uw. In *Spring 2003 EARS workshop.* Boston, MA.

Tseng, C.-Y., Pin, S.-H., Lee, Y., Wang, H.-M., & Chen, Y.-C. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication*, *46*, 284-309.

Turk, A. E., & Sawusch, J. R. (1997). The domain of accentual lengthening in american english. *J. Phonetics*, *25*, 25-41.

van Kuijk, D., & Boves, L. (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication*, *27*, 95-111.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707-1717.

Woodland, P., Hain, T., Evermann, G., & Povey, D. (2001). Cu-htk march 2001 hub5 system. In *Nist 2001 workshop on speech transcription.*

Yoon, T. (2007). *A predictive model of prosody through grammatical interface: A computational approach.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using tobi. In *Proc. internat. conf. spoken language processing.* Jeju Island, Korea.

Yoon, T., Zhuang, X., Cole, J., & Hasegawa-Johnson, M. (2006). Voice-quality dependent automatic speech recognition. In *Linguistic patterns in spontaneous speech.* Taipei: Academica Sinica.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2002). *The HTK book.* Cambridge, UK: Cambridge University Engineering Department.

Zhang, Y. (2000). *Information fusion for robust audio-visual speech recognition.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Zheng, Y., & Hasegawa-Johnson, M. (2004). Stop consonant classification by dynamic formant trajectory. In *Proc. interspeech.* Jeju Island, Korea.

Zue, V., & Laferriere, M. (1979). Acoustic study of medial /t,d/ in american english. *Journal of the Acoustical Society of America*, *66*(4), 1039-1050.

Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, *9*, 351-356.

Zweig, G., Bilmes, J., Filali, K., Livescu, K., Xu, P., Jackson, K., Brandman, Y., Sandness, E., Holtz, E., Torres, J., & Byrne, B. (2002). Structurally discriminative graphical models for automatic speech recognition—results from the 2001 johns hopkins summer workshop. In *Proc. icassp* (p. 183-190). Denver, CO.