Sub-word modeling for automatic speech recognition

Karen Livescu, *Member, IEEE*, Eric Fosler-Lussier, *Senior Member, IEEE*, Florian Metze, *Member, IEEE*

***Abstract*—Modern automatic speech recognition systems handle large vocabularies of words, making it infeasible to collect enough repetitions of each word to train individual word models. Instead, large-vocabulary recognizers represent each word in terms of sub-word units. Typically the sub-word unit is the phone, a basic speech sound such as a single consonant or vowel. Each word is then represented as a sequence, or several alternative sequences, of phones specified in a pronunciation dictionary. Other choices of sub-word units have been studied as well. The choice of sub-word units, and the way in which the recognizer represents words in terms of combinations of those units, is the problem of *sub-word modeling*. Different sub-word models may be preferable in different settings, such as high-variability conversational speech, high-noise conditions, low-resource settings, or multilingual speech recognition. This article reviews past, present, and emerging approaches to sub-word modeling. In order to make clean comparisons between many approaches, the review uses the unifying language of graphical models.**

I. Introduction

Automatic speech recognition has enjoyed decades of progress, including the successful introduction of commercial voice-based services. However, there are still unsettled questions in the speech recognition research community, and one of these is how to model the internal structure of words. The main questions are *what are the basic units that should be modeled?* and *how should the structure over these units be modeled, parameterized, and trained?* This article gives an overview of potential answers to these questions, including a historical overview, a description of the current state of this research area, and presentation of emerging techniques that may affect the future state of the art.

Throughout the article, we assume that the task of interest is word recognition. That is, given an acoustic recording of a sequence of one or more spoken words, the task is to infer the word(s). We implicitly assume that the language is known but not necessarily that the speaker identity is known (i.e., we consider speaker-independent recognition). We begin in this section by setting the stage: why sub-word units are needed, what the most common sub-word models are, and why alternatives have been considered.

A. Why sub-word models?

Why should words be broken up into smaller units at all? The word recognition problem could be framed as a comparison between a test pattern—typically a spectral representation of an input waveform—and stored reference patterns for words. In order to account for variations in production, we should have many stored examples of each word.

For any recognition task with a large vocabulary, the whole-word approach is impractical. Words are distributed approximately according to Zipf's law, i.e. the frequency of a word is roughly inversely proportional to its rank in the frequency table. In the 3-million-word Switchboard-1 Corpus of telephone conversations, the 43 most frequent word types account for half of the word tokens, while the other half are distributed across about 33,000 word types. Some words—many names, new coinages, words related to current events—may not occur at all in any finite corpus of recorded speech. Unfortunately, these words are often relevant in practice.

This observation motivates the use of sub-word units that occur often in reasonably sized speech corpora. If we have no recordings of, say, the word "batrachophagous", we may hypothesize that it starts with the same consonant sound as "bar", continues with the same vowel sound as in "hat", and so on. If we have sufficiently many recordings of these individual sounds, so-called *phones*, perhaps we can build a model of the word out of models of the phones, and collect pronunciations in a phonetic dictionary. Alternatively, we could note that the word starts with the same entire first syllable as "batter" does, ends with the same syllable as "analogous", and so on. If we have sufficiently many recordings of all possible syllables, we can then build word models by concatenating syllable models.

Phones and syllables, then, are potential types of sub-word units. Additional alternatives are *sub-phonetic features* ("batrachophagous" starts with a *stop consonant*, which is also *voiced* and produced at the *lips*); *graphemes* ("batrachophagous" starts with the same two letters as "bar", so it might also start with similar sounds); and *automatically learned sub-word units*,

which are units corresponding to acoustic segments that have been consistently observed in training data.

Good sub-word units should be (1) *trainable*; i.e., they should be sufficiently frequent in typical corpora, (2) *generalizable*; i.e., they should be able to represent previously unseen words during testing, and (3) *invariant*; i.e., they should be robust to changes in environment and context. Choosing the best type of unit, and the best associated model of word structure, is a critical decision point in the speech recognition problem: Virtually all other components and algorithms in speech recognition presuppose the existence of a fixed set of units.

B. Phones and context-dependent phones

The most commonly used sub-word units are phones.[1] There are typically 30-80 phones per language. In today's recognizers, words are usually represented as one or more phone sequences, often referred to as the "beads-on-a-string" representation [1], [2]. Common variants may be listed ("going" → [g ow ih ng], [g ow ih n]) or generated by rule ("-ing" → [ih ng], [ih n]).[2]

The same phone may be realized differently in different contexts, due to coarticulation, stress, and other factors. For example, [k] is usually articulated with the tongue farther forward in the word "keep" and farther back in the word "coop", resulting in very different signals. To take such effects into account, each phone in each relevant context can be considered a separate unit. This is the *context-dependent phone* unit used in most speech recognizers [3]. Automatically learned decision trees are used to partition the data into roughly homogeneous acoustic units, usually based on the preceding and following phones; depending on the context window size, the resulting units are called *triphones* (for a ± 1 phone context), *quinphones* (for ± 2 phones), and so on. Each context-dependent unit is typically represented by a hidden Markov model (HMM) with Gaussian mixture observation densities, which account for the remaining acoustic variation among different instances of the same unit. For further details about the architecture of standard HMM-based recognizers, see [4].

[1]Linguists distinguish phones—acoustic realizations of speech sounds—from *phonemes*—abstract sound units, each possibly corresponding to multiple phones, such that a change in a single phoneme can change a word's identity. In speech recognition research, these terms are often used interchangeably, and recognition dictionaries often include a mix of phones and phonemes. We will use the term "phone" throughout as it is more typical in speech recognition, although we will distinguish between canonical phones (found in a dictionary) and surface phones (that are observed). The entire discussion applies similarly to phones and phonemes.

[2]We use the ARPA phonetic alphabet for English examples.

C. Challenges for sub-word models

Given the above, why is sub-word modeling not a "closed case"? Two main challenges dominate the discussion: pronunciation variability and data sparseness.

a) Pronunciation variability: Spoken words, especially in conversational speech, are often pronounced differently from their dictionary pronunciations (also referred to as canonical pronunciations or baseforms) [5], [6]. This variability is the result of many factors—the degree of formality of the situation, the familiarity of speakers with their conversation partners and relative seniority, the (presumed) language competency of the listener, and the background noise [7]—and is one of the main challenges facing speech recognition [8]. Context-dependent phones and Gaussian mixtures cover a great deal of the variation, in particular substitutions of one sound for another; but some common pronunciation phenomena, such as apparent deletions of sounds, are poorly accounted for [9]. The performance of speech recognizers degrades sharply on conversational speech relative to read speech, even when exactly the same word sequences are spoken by the same speakers in the same acoustic environment; in other words, conversational pronunciation style alone is responsible for large performance losses [5]. Even within a single sentence, different words may be pronounced more or less canonically, and the ones that are pronounced non-canonically tend to be misrecognized more often [10].

Perhaps more surprisingly, "hyper-clear," or over-emphasized, speech degrades recognition performance as well [11], although it can improve intelligibility for humans [12]. That speech recognition is worse for both conversational and "hyper-clear" speech suggests that the representations used in today's recognizers may still be flawed, despite impressive progress made over the years.

Figure 1 shows an example of the types of variation seen in the Switchboard conversational speech corpus, as transcribed by expert phonetic transcribers [13]. Other examples from the same corpus include the word "probably" with such pronunciations as [p r aa b iy], [p r ay], and [p r aw l uh], and "everybody" with pronunciations such as [eh v er b ah d iy], [eh b ah iy], and [eh r uw ay]. Overall, fewer than half of the word tokens in this corpus are pronounced canonically.

This variability is not language-specific. In German, "haben wir" ("we have") is canonically pronounced [h a: b @ n v i:6] (using the SAMPA international transcription alphabet), but can be pronounced as [h a m a] or [h a m v a] in colloquial speech. Similar examples occur in French; e.g., *cinéma*: [s i n e m a] → [s i n m a], *c'est pas*: [s E p a] → [s p a] [14].

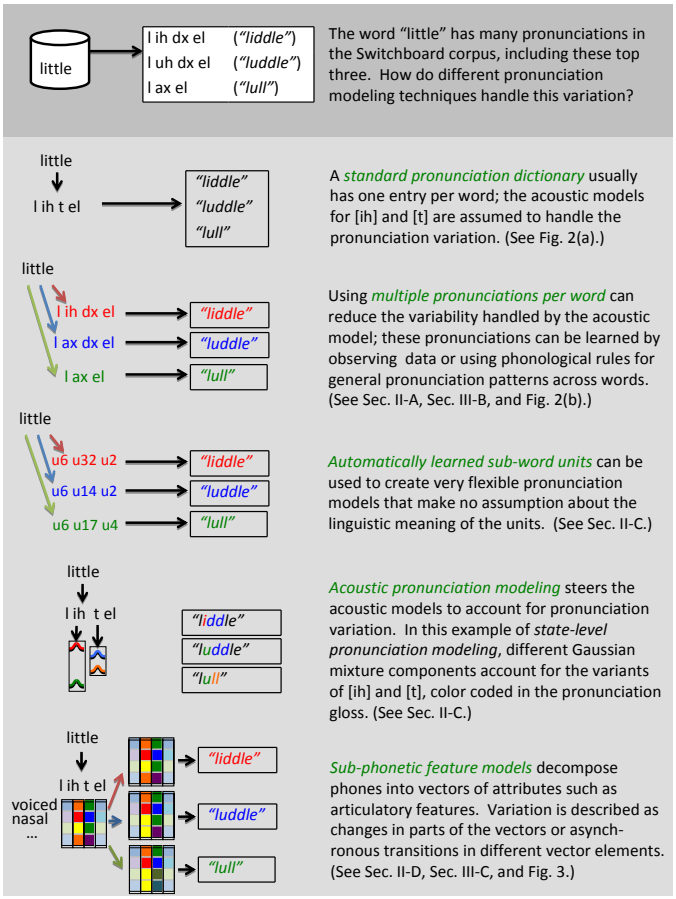


Fig. 1. An example of pronunciation variation in conversational speech, a standard dictionary representation, and four alternative approaches to describing this variation, serving as an informal description of techniques described in Sections II–III.

However, pronunciation changes do not necessarily occur at the level of entire phones. Instead, changes often occur at a sub-phonetic level, such as devoicing of voiced consonants, spreading of rounding to phones near a rounded phone, or nasalization of vowels near nasal consonants.

b) Data sparseness: Another challenge is the number of sub-word units relative to the amount of training data available. For example, there are tens of thousands of triphone units that occur in a typical language. This makes it difficult to train conventional models for languages or dialects in which few resources (audio data, dictionaries) are available. A recent interest in open-vocabulary and spoken term detection systems, in which the vocabulary (or even the precise dialect or language) may not be known in advance, creates an additional incentive to investigate models based on units that are more language-independent and robust to data sparseness. Such considerations have also motivated approaches using smaller inventories of "universal" sub-phonetic units that can be combined in many ways to form the sounds of the world's languages and dialects.

These two challenges—pronunciation variability and data sparseness—contribute to keeping speech recognition from being used for unrestricted applications, such as court room transcription, closed captioning, free-style dialogue systems, and quickly portable cross-language applications. For example, large-vocabulary English conversational telephone speech is currently recognized at roughly 20% word error rate; this is sufficient for some tasks, like searching for content words, but does not make for a reliable, readable transcript.

Besides these two challenges, some researchers feel that speech recognition in general would be improved by using sub-word models that are more faithful to knowledge from linguistics and speech science. This consideration has also motivated some of the approaches described here, although we focus on the motivations presented by the pronunciation variation and data challenges, and will not comment on the fidelity of the approaches to human speech processing.

II. Historical Review

Since the first large-vocabulary speech recognition systems of the mid-1970s, the predominant type of sub-word model has been the representation of a word as one or more strings of phones [15]. Throughout the intervening years, however, a variety of alternative sub-word models have been studied in parallel, with the basic units including syllables [16], [17], acoustically defined units [18], [19], graphemes [20], and sub-phonetic features [21], [22], [23], [24], [25], [26]. Figure 1 serves as an informal summary of some of the main sub-word modeling approaches described in this article.

A. Dictionary expansion

In the 1990s and early 2000s, interest in conversational speech recognition led to several studies on the properties of the conversational style and its effects on recognition performance [5], [10], [7], [13], [9], [2]. This led to a great deal of activity on modeling pronunciation variation, including two workshops sponsored by the International Speech Communication Association [27], [28]. The majority (but by no means all) of the proposed approaches during this period kept the phone as the basic sub-word unit, and focused on ways of predicting the possible phonetic sequences for any given word using phonological rules or other means [29], [30], [31], [32], [33], [7], [34], [10].

In the example of Figure 1, the first vowel of the word "little" exhibits variation from the expected [ih] phone to the sounds [ax] and [uh] that are produced farther back in the mouth. In addition, in American English, the "t"

sound in this context is usually realized as the flap [dx] (with the tongue briefly touching the roof of the mouth), but it can also be deleted.

One way to account for such variation is to include all observed variants in the pronunciation dictionary, perhaps along with the probability of seeing each variant. A large amount of work in sub-word modeling has involved such expansion of the baseform dictionary with additional pronunciations [31], [7], [35]. However, when pronunciation variants are learned separately for each word, frequently observed words may develop a rich inventory of variants, while infrequent words may be poorly modeled: Unless some capability for generalization is built in, learning new variants for *little* will not inform about variants for *whittle* and *spittle*.

One common approach to increase generalization is to model pronunciation changes as transformations from one phone sequence (a canonical pronunciation) to another (an observed surface pronunciation) via phonological rules. A phonological rule can be represented as a transduction from a string of phones X to a string of phones Y when surrounded by a particular context. For example, the flapping of [t] in *little* could be generated by the rule $\{$ ih **t** ax $\rightarrow$ dx $\}$ (read "[t] can be realized as a flap between the vowels [ih] and [ax]"). Such rules can be specified manually from linguistic knowledge [33] or learned automatically, typically using decision trees [30], [10]. Once a set of rules is specified or learned, it can be used to expand a dictionary to form a single new dictionary (a *static* expansion) or to expand the dictionary *dynamically* during recognition in response to unpredictable context such as the speaking rate [10].

Learning the distribution over phonetic baseforms or rules requires phonetically labeled training data. This can be obtained from manual transcriptions [30] or using a phonetic recognizer. The learning has typically been done by maximizing the model likelihood over the training data [31], although in some work a discriminative objective is optimized instead [35], [32].

B. Impact of phonetic dictionary expansion

Phonetic dictionary expansion has produced improvements in some systems [30], [10], [33]. However, the improvements have been more modest than hoped, considering the very large difference in performance on read and conversational renditions of the same word sequences [5]. One issue is the tradeoff between *coverage* and *confusability*. As pronunciations are added to a dictionary, coverage of alternative pronunciations is improved, while at the same time, words become more confusable due to increasing overlap in their allowed pronunciations. Several researchers have tried to quantify confusability [36], [37], to limit the amount of variation to just the minimum needed [38], or to use discriminative training to eliminate error-causing confusions [35], but balancing confusability and coverage remains an active area of research.

Reference	Type of error	Error rate (%)
Canonical	ERR	14.2
	DEL	2.4
	SUB	11.8
Non-canonical	ERR	20.7
	DEL	4.2
	SUB	16.5
None	INS	10.8
All	ERR	27.2

TABLE I
RECOGNITION RESULTS (ERROR PERCENTAGE IS THE SUM OF DELETIONS AND SUBSTITUTIONS) FOR WORDS PRONOUNCED CANONICALLY AND NON-CANONICALLY. BOTTOM TWO ROWS: OVERALL INSERTION RATES AND OVERALL WORD ERROR RATES.

The current mainstream approach—that is, the approach typically used in state-of-the-art systems in benchmark competitions—uses a phonetic baseform dictionary with a single pronunciation for most words, a small number of variants for remaining, frequent words. Dictionaries are typically manually generated, but can also be generated in a data-driven way [39], [40].

To show the influence that pronunciation variation still has on today's systems, we analyze the hypotheses of a state-of-the-art conversational speech recognizer, similarly to earlier experiments described in [10]. Table I shows the results of testing the recognizer on phonetically transcribed data from the Switchboard Transcription Project [13].[3] Approximately 60% of the word tokens in the test set were manually labeled as having a non-canonical pronunciation. There is approximately a 50% increase in errors when a word is pronounced non-canonically, similarly to earlier findings [10].

We now continue our historical review with some alternatives to phonetic dictionary expansion.

C. Acoustics-based models

Investigations of the acoustic realizations of phones labeled by linguists as non-canonical have shown that the acoustics are often closer to the canonical phone than

[3]Some technical details: The recognizer is speaker-independent, without adaptation but with discriminative training using a maximum mutual information criterion on a standard 350-hour training set [41], using a trigram language model and a vocabulary of 50,000 words. The dictionary contains multiple pronunciations for a subset of the words, for a total of 95,000 variants, derived from the CMU dictionary using knowledge-based phonological rules. Acoustic model training uses Viterbi alignment of multiple pronunciation variants.

to the putative transcribed phone [42], so a replacement of an entire phone in the dictionary may be an inaccurate representation of the change. Given this continuous nature of pronunciation variation, combined with the limited improvements seen from dictionary expansion, some proposed that the handling of pronunciation variation may be better done using acoustically defined units [43], [18] or by modifying the acoustic model of a phone-based recognizer [44], [45].

In *acoustically defined sub-word unit models*, an alternative to the phone is sought that better describes speech sound segments. One typical approach [43], [18] is to first segment observed word tokens into a number of coherent regions, then cluster these regions to produce a set of units that generalize across the words in the vocabulary (like the numbered units, e.g. *u6*, in Figure 1). The appeal of such an approach is that, since the pronunciations are derived directly from audio, it should be better tuned to the task than a dictionary model. However, there are a few challenges as well: Deriving representations for words not seen in training is difficult since the usual prior mapping from words to sounds is unavailable; building context-dependent acoustic models is also problematic, as the typical decision tree clustering algorithms ask questions about the *linguistic* nature of neighboring units, which is unavailable here. Another active research direction is the use of alternative modeling techniques that do not follow the segment-then-cluster approach [46], [19], [47].

Acoustic pronunciation modeling, in contrast, uses modified acoustic models combined with a basic phone-based dictionary. One such strategy is *state-level pronunciation modeling* [45]. This technique starts with standard mixture of Gaussian observation models trained using a canonical pronunciation dictionary, and then combines Gaussians from phones that are found to be frequent variants of each other in phonetic transcriptions. For example, in Figure 1, the pronunciation of the vowel in *little* may borrow Gaussians from both the [ih] model and the [ax] model seen in one of the variants.

A similar intuition led to the *hidden model sequence HMM (HMS-HMM)* approach proposed by Hain [48], [49], in which each phone is represented by a mixture of HMM state sequences corresponding to different variants. Both state-level pronunciation modeling and hidden model sequences, then, account for the continuous nature of pronunciation variation by making "soft" decisions about phone changes. Hain also proposed a procedure for iteratively collapsing multiple dictionary pronunciations to a single pronunciation per word, based on observed frequencies in training data, and extrapolating to unseen words, which produced the same performance on conversational speech as the original multi-pronunciation dictionaries [44], [49]. Such a procedure tunes lexical representations to the acoustic models that are accounting for some of the phonetic variation. Nevertheless, most state-of-the art systems use dictionaries with multiple variants for frequent words with variable pronunciation, rather than tuning a single-pronunciation dictionary to a specific data set and acoustic model.

D. Sub-phonetic feature models

One of the primary differences between explicit phone-based models and acoustics-based models is the granularity: Phone-based models describe variation as discrete changes in phonetic symbols, but may not capture subtle acoustic variation; acoustics-based models give a fine-grained, continuous view of pronunciation variation, but may miss opportunities for generalization. A middle ground is to *factor the phonetic space* into sub-phonetic feature units. Typical sub-phonetic features are *articulatory features*, which may be binary or multi-valued and characterize in some way the configuration of the vocal tract.[4] Roughly 80% of phonetic substitutions of consonants in the Switchboard Transcription Project data consist of a single articulatory feature change [10]. In addition, effects such as nasalization, rounding, and stop consonant epenthesis can be the result of asynchrony between articulatory trajectories [50]. A factored representation may allow for a more precise and parsimonious explanation of these phenomena. In addition, such a representation may allow for reuse of data across languages that share sub-phonetic features but not phone sets, thus helping in multilingual or low-resource language settings [51].[5]

Two general approaches have been used for sub-phonetic modeling in speech recognition. The first is what we refer to as *factored observation models* [26], [53], [25], [54], where a standard phonetic dictionary is used, but the acoustic model consists of a product of distributions or scores, one per sub-phonetic feature, and possibly also a standard phonetic acoustic model. Factored observation models address the challenge of robustness in the face of data sparseness, and may also be more robust to noise [26]. They do not, however,

[4]We use the term "articulatory features" to refer to both discretized positions of speech articulators and more perception-based *phonological features* such as manner and place of articulation. These terms are sometimes distinguished, but not consistently so in the speech recognition literature. For this reason we use a single term for both.

[5]Earlier, related work used state-based tying and multilingual phone sets to learn which phones to share and which phones to separate between languages automatically from data [52] in low-resource and multilingual settings.

explicitly account for changes in articulatory feature values or asynchrony. To address this, some have proposed representing the dictionary explicitly in terms of sub-phonetic features. In this approach, sometimes inspired by the theory of articulatory phonology [55], many effects in pronunciation variation are described as the result of articulatory asynchrony and/or individual feature changes. We refer to this as a *factored-state* approach because the hidden phonetic state is factored into multiple streams.

One of the first series of investigations into a factored-state approach was by Deng and colleagues [21], [56], [57], using HMMs similar to those of standard phone-based models, but with each state corresponding to a vector of articulatory feature values. All possible value combinations are possible states, but transitions are constrained to allow only a certain amount of asynchrony. More recently, a more general approach to articulatory pronunciation modeling has been formulated, in which graphical models represent both articulatory asynchrony and deviations from articulatory targets [58], [59], [60]. In factored models using articulatory features, it is possible to use articulatory inversion as a form of observation modeling [61], or to use generative observation models [21], [59] (see, e.g., [24] for a review of techniques). Here, however, we restrict our attention to the mapping between words and sub-word units.

E. Conditional models

In the approaches described thus far, each word consists of some combination of sub-word units that must be present. Another recent line of work involves conditional models (also referred to as direct models [62]), which changes the nature of the relationship between words and sub-word units. In this approach, sub-word representations are thought of as evidence of the presence of the word [63], [64], [65]. In contrast to generative models like HMMs, conditional models directly represent posterior probabilities, or more generally scores, of the unknown labels (words, states) given observations (acoustics) and are trained by optimizing criteria more closely related to the prediction task. Such approaches have been developed as extensions of conditional models for phonetic recognition [66], [67], but they serve as new forms of sub-word modeling in their own right (although they are not necessarily framed in this way). The conditional approach allows for multiple, overlapping sub-word representations that can be combined in ways that are difficult to do in traditional HMM-based models [65].

III. Sub-word models as graphical models

Many of the approaches reviewed above fit into the standard speech recognition framework of HMM-based modeling, but some do not. The development of some of the sub-phonetic and conditional models discussed above has been facilitated by the rise of graphical model techniques, which generalize HMMs and other sequence models. Graphical models have been gaining popularity in speech recognition research since the late 1990s, when dynamic Bayesian networks (DBNs) were first used to represent HMM-based speech recognizers and then to introduce additional structure [68], [69], [70]. In order to easily compare various approaches, this section unifies much of the prior and current work on sub-word modeling in a graphical model representation. We first define graphical models, and then formulate several types of sub-word models in this representation.

A. Brief introduction to graphical models

A graphical model [71] is a representation of a probability distribution over N variables $X_1, \ldots, X_N$ via a graph, in which each node is associated with a variable X_i. The graph encodes the factorization of the distribution as a product of functions, each of which depends on only a subset of the variables. Graphical models have become a *lingua franca* of machine learning and artificial intelligence [72], because they can parsimoniously represent complex models and because there are uniform algorithms for doing computations with large classes of graphical models. The main type of computation is inference—"given the values of the variables in set A, what is the distribution (or most probable values) of the variables in set B?"—which is required for both testing (doing prediction with) and training (learning parameters for) a graphical model.

In *directed* graphical models, or *Bayesian networks (BNs)*, the joint distribution is given by the product of the "local" conditional distributions of each variable X_i given its parents in the graph $pa(X_i)$:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | pa(x_i)). \quad (1)$$

We use lower-case letters to denote the values of random variables, e.g. x is a value of X. A *dynamic Bayesian network (DBN)* consists of repeating sub-graphs, or *frames*. DBNs are appropriate for modeling stochastic processes over time, such as speech (where the frame may correspond to the usual 10ms frame of speech). An HMM is a special case of a DBN in which each frame consists of a state variable and an observation variable.

Conditional random fields (CRFs) [73] are undirected models of a *conditional* distribution $p(Q|O)$. Given the observed variables $O = (O_1, \ldots, O_L)$, the joint distribution over the hidden variables $Q = (Q_1, \ldots, Q_M)$

is given by the product of local *potential functions* $\psi_k\left(Q_{\{k\}}, O\right)$ over cliques of variables $Q_{\{k\}}$, $k \in \{1, \ldots, K\}$:

$$p(q_1, \ldots, q_M|o) = \frac{1}{Z(o)} \prod_{k=1}^{K} \psi_k\left(q_{\{k\}}, o\right), \qquad (2)$$

where $Z(o)$ is a normalizing constant. The potential functions are typically assumed to have a log-linear form $\psi_k\left(q_{\{k\}}, o\right) = \exp\left(\sum_j \theta_{kj} f_{kj}\left(q_{\{k\}}, o\right)\right)$, where the *feature functions* f_{kj} are fixed and only the weights θ_{kj} are learned.[6] This means that the predictor function for the hidden variables, $\arg\max_{q_1,\ldots,q_M} p(q_1, \ldots, q_M|o)$, has the form of a summation over the weighted feature functions, similarly to other discriminative models like structured SVMs [74]. A recent variant is *segmental CRFs* [65], in which each hidden variable may extend over a varying number of frames.

We do not address the important problem of effective and efficient inference for different types of models; the reader is referred to previous review articles and texts [70], [72]. The parameters of generative graphical models can be learned either with the classic expectation-maximization (EM) algorithm [75] or with discriminative training algorithms (e.g., [76]). For discriminative models, a number of learning approaches such as maximum conditional likelihood (as in CRFs [73]) or large-margin training (as in structured SVMs [74]) are used.

Directed models are particularly useful when interpretability is important. Undirected models are useful for combining many different information sources (via the feature functions).

B. Phone-based models as DBNs

Figure 2 shows several phone-based models represented as DBNs (although they are not typically implemented as DBNs). Figure 2(a) represents a standard HMM-based speech recognizer with a single baseform pronunciation per word. This DBN is simply an encoding of a typical HMM-based recognizer. Without loss of generality, we refer to the sub-word variable q_t as the phone state; however, this variable may represent either a sub-phonetic monophone state ([ih1], [ih2], [ih3]) or a context-dependent phone (e.g., triphone) state.

The DBN of Figure 2(a) is a complete speech recognizer, except for certain details of the language model. The sub-word model is that portion that concerns the mapping between words and phone states. In the remaining models below, we will only present the variables and dependencies involved in sub-word modeling; that is, we will not show the word and observation variables.

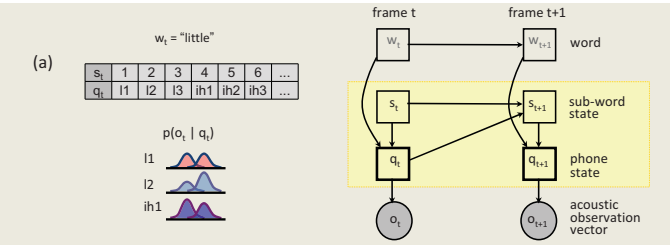
The remainder of Figure 2 shows alternative phone-based sub-word models. Figure 2(b) shows a sub-word model with multiple pronunciations per word—which represents, more or less, the mainstream approach—and Figure 2(c) shows a model in which the multiple pronunciations are generated by applying context-dependent probabilistic phonological rules represented as decision trees, which involves adding variables to the DBN corresponding to the desired context. In Figure 2(c), the context variables are deterministic given the sub-word state (e.g., properties of the previous and next phones). In general, the context variables may be more complex—e.g., higher-level context such as word frequency or speaking rate—which may require different dependencies. The distribution of the context-dependent phone state variable $p(q_t|u_t, c_t^1, c_t^2, \ldots)$ is typically not learned jointly with the other parameters, but rather decision trees are separately learned for predicting the phone distribution given the context variables [30], [10]. In other work, rule "firing" probabilities are learned separately or as part of the complete recognizer [33]. In addition, the same model structure can describe certain acoustics-based models; for example, Hain's HMS-HMM approach (Section II-C) [48] has the same structure except that the "surface phone state" is an abstract HMM model state, and is not shared across canonical phones with similar surface realizations.

C. Sub-phonetic feature models as DBNs

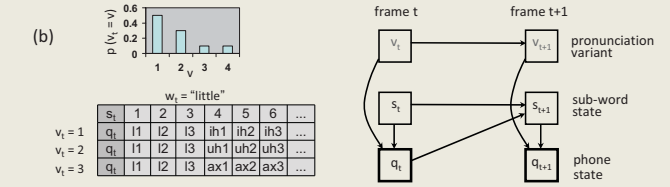
Figure 3 shows sub-phonetic feature models (see Sec. II-D) represented as DBNs. Figure 3(a) represents *factored observation models*, in which the phone state variable q_t is mapped to multiple feature state variables q_t^i, each of which is associated with a separate observation distribution $p(o_t|q_t^i)$ (e.g., Gaussian mixtures as in [25], [23]) or separate discriminative classifier [26], [77] for each sub-phonetic feature i, and optionally an additional standard observation distribution per phone state $p(o_t|q_t)$. If classifiers are used, their outputs are either scaled to produce scaled likelihoods $\propto p(o_t|q_t^i)$ [26] or used as new observation vectors over which Gaussian mixture distributions are trained [77]. These distributions/scaled likelihoods are multiplied to produce the final observation model.

Figure 3(b) shows a factored-state model, with no phone state variable at all, based on [60]. Each sub-phonetic feature follows its own trajectory through the state sequence of each word. In this case the feature streams correspond to specific articulators such as the

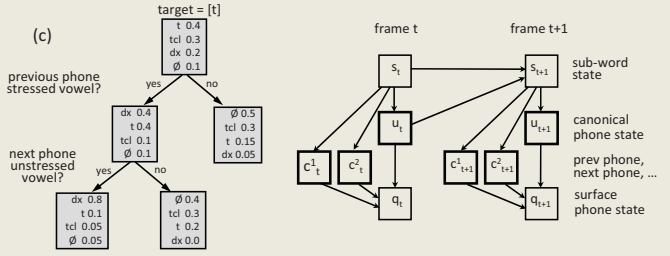
[6] Feature functions are sometimes also referred to as features. We use the term feature *functions* throughout to avoid confusion with sub-phonetic features.



A *standard phone-based speech recognizer*. The portion corresponding to the sub-word model is boxed in yellow; subsequent figures include only the sub-word model. The sub-word state variable steps through the integers 1, 2, …, L where L is the number of states in the word's canonical pronunciation. In each frame, the state may stay constant or increment, depending on the transition probability of the current phonetic state q_t. To the left are depictions of the distributions of the phone state and observation vector variables. The phone is given deterministically by the current word and sub-word state. The observation vector has a Gaussian mixture distribution conditioned on the phone state.

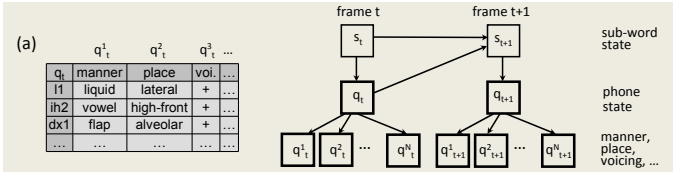


A *model with a multiple-pronunciation dictionary*. Word and observation variables, and edges to/from them, have been omitted. This model differs from the baseline above in the addition of one variable, the pronunciation variant, which is an additional parent to the phone state variable and stays constant within a word. To the left is an example distribution of the pronunciation variant variable, as well as a table mapping from the word, pronunciation variant, and sub-word state variables to the phone state.

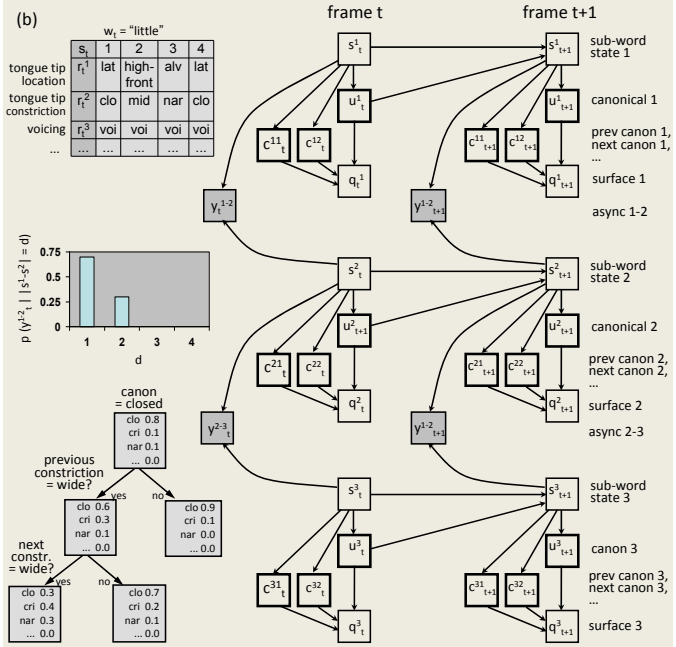


A *sub-word model with context-dependent phonological rules*. This model differentiates between the canonical phone in the dictionary and the surface phone produced, depending on context variables such as the next or previous phone. The distribution of the surface phone is often modeled using a decision tree; an example tree for (all states of) the target phone [t] is shown at left.

Fig. 2. Phone-based sub-word models as DBNs. Notation: square/circular nodes correspond to discrete/continuous variables; shaded nodes are observed; nodes with thick outlines are deterministic given their parents. Here and throughout, we omit certain details, such as the special cases of the initial and final frames, distinctions between training and decoding models, and precise representation of the language model (in fact the above is a precise representation of an isolated-word recognizer). See [70] for more information about DBNs for speech recognition.



A *factored observation model* where the factors correspond to sub-phonetic features. The table at left shows a portion of a possible mapping from phones to articulatory features. When using such a model in a complete recognizer, the observation vector depends on the articulatory features, as well as possibly the phone.



A *factored-state model*, which models observed pronunciations as the result of asynchrony between articulators and substitutions in each articulatory feature stream. Here the sub-word units that the acoustics ultimately depend on are the actual articulatory feature values. Each articulatory feature stream is associated with its own sub-word state variable; given the word and sub-word state, the target positions of the articulators are given by the dictionary pronunciation(s) shown in the table on the left. If the articulators move asynchronously, those sub-word state variables have different values. The amount of asynchrony is controlled by asynchrony variables $y^{i\text{-}j}_t$, accounting for effects like anticipatory nasalization, rounding, and epenthetic stops. The articulators may also fail to reach their target positions. The distribution of actual articulatory positions, given the targets and possibly other context variables, is here described by a decision tree, similarly to context-dependent phone-based models.

Fig. 3. Sub-word models based on sub-phonetic features.

lips, tongue, glottis, and velum. Note that the sub-word sub-structure for each feature is analogous to the structure of the phone-based model of Figure 2(c). As in phone-based models, context-dependent deviations from canonical values can be modeled using decision trees. Note the similarity between the structure in Figure 2(c) and the feature-specific substructures in Figure 3(b). In Figure 3(b), each surface feature value depends on its canonical target as well as the previous and next canonical targets, which takes into account the tendency of articulators to assimilate with their past/future states. Many additional context variables are possible [60]. Since the features now each have their own sub-word state variable, they may not proceed through the word

synchronously. The model probabilistically constrains the asynchrony between features via the asynchrony variables. Such models have a fairly complex graphical structure, but by virtue of factoring the state distribution, they can have fewer parameters than analogous phone-based models (Fig. 2) and than factored-state models represented as HMMs [21], [22].

D. Conditional models

As mentioned in Section II, conditional models are becoming increasingly popular for representing various aspects of speech recognition, including sub-word models. In terms of their graphical model structure, the models that have been developed thus far are essentially the undirected equivalents of the models in Figures 2 and 3. The key point in these models is how the feature functions over cliques of variables are defined.

1) Phone-based CRFs: The most commonly used conditional models are conditional random fields (CRFs), defined above in Equation 2. Analogues of the basic phone-based model of Figure 2(a) have been investigated extensively for phonetic recognition [66], [67], [54]. A direct analogue of a single-Gaussian HMM corresponds to using the Gaussian sufficient statistics (the acoustic observations and their inner products) as feature functions. If a hidden state variable is also added [66], [67], the model becomes analogous to a HMM with a Gaussian mixture acoustic model. The key differences are the conditional training and the ability to include additional feature functions.

Different choices of feature functions can give rise to different types of models; for example, using posteriors over sub-phonetic feature classes as feature functions results in a system that is analogous to the factored observation model of Figure 3(a) [54]. CRF models with similar structure to the articulatory DBN of Figure 3 have also recently been introduced [78].

Segmental CRFs (SCRFs) have also been used as a form of sub-word modeling. For example, in [65], the authors define SCRF feature functions that correspond to aligned pairs of expected phone sequences and observed ones, which is the analogue of context-dependent phonological rules in prior phone-based work (Sec. II-A). They also use additional new feature functions, such as co-occurrence (without an explicit alignment) of baseform phone sequences and surface phone sequences. This framework allows for a very rich set of feature functions, since any functions spanning a word unit can be used. Models of the same form as SCRFs can in principle be trained with other discriminative criteria and features functions, as done in [79] with large-margin training and feature functions combining phone-based and articulatory information.

Approach	Result
Decision tree-based phononological rules [30] *Figure 2(c)*	Improvements over baseline dictionary by 1-3% on conversational speech and broadcast news recognition.
Dynamic phonological rules using phonetic, prosodic, etc. context [10] *Figure 2(c)*	Improvements over baseline dictionary by 3-5% on conversational speech.
Segment-based system with phonological rules [33] *Figure 2(c)*	Improvement over baseline dictionary by 9% on medium-vocabulary weather query task.
Discriminative selection of pronunciation variants [35] *Figure 2(b)*	Improvement over baseline dictionary by 7% on recognition for voice search.
Automatically learned sub-word units [18], [80] *Figure 2(a)*	Allows automatically inducing dictionary from data; 3% improvement over phonetic baseline system for conversational speech, larger improvements on small-vocabulary task.
State-level pronunciation modeling [42] *Figure 2(a)*	Improvement over standard HMMs by 5% on conversational and read speech.
Hidden model sequences [49] *Figure 2(a)*	Improvement by up to 4% over standard HMMs on conversational telephone speech.
Factored articulatory observation model using multilayer perceptrons [26] *Figure 3(a)*	Improvement of $\sim$5% over unfactored phone-based model in noisy medium-vocabulary speech recognition.
Factored articulatory observation model using Gaussian mixtures [25], [11], [51], [23] *Figure 3(a)*	Improvements on large-vocabulary & cross-lingual recognition, hyper-articulated speech recognition, and small-vocabulary recognition in noise by 5-10%.
Factored-state model using articulatory features [22] *Figure 3(b)*	Improvement of $\sim$25% in combination with a baseline HMM on medium-vocabulary isolated words.
Segmental CRFs with phone-based feature functions [65]	Improvement of $\sim$10% over state-of-the-art baseline generative model on broadcast news recognition.

TABLE II
SAMPLE OF RESULTS FROM THE LITERATURE ON SUB-WORD MODELS IN SPEECH RECOGNITION. ALL NUMERICAL RESULTS REFER TO RELATIVE IMPROVEMENTS (E.G., ERROR RATE REDUCTION FROM 20% TO 18% IS A 10% IMPROVEMENT).

IV. EMPIRICAL MODEL COMPARISONS

To give an idea of the current state of sub-word modeling research, we provide selected results from the literature in Table II. A head-to-head comparison has not been done for most of the models discussed here, so reported performance improvements are specific to a particular type of recognition system, task (e.g., larger vs. smaller vocabulary), and choice of data. We provide

a sample of the reported results in terms of relative improvement, the percentage of errors made by a baseline system corrected by a proposed approach. Some of the approaches have been applied to phonetic recognition; here we include only word recognition results. The first four lines in the table describe phone-based dictionary expansion techniques discussed in Section II-A. The next three lines refer to acoustics-based approaches (Section II-C). Here the goals of the approaches differ somewhat: While all aim to improve recognition performance, automatically learned units also allow learning the pronunciation dictionary from data. The next three lines give results of sub-phonetic feature-based models (Section II-D). While these have shown some gains in performance, they have largely not yet been incorporated into large-scale state-of-the-art systems. Finally, the last line gives an example of a conditional model (Section II-E) with feature functions encoding sub-word structure. While many of the approaches show significant improvement over single-/multiple-pronunciation phone-based systems, at least 75% of the errors are not corrected by any of the approaches, leaving this area still open for wide-ranging research.

It is sometimes useful to test a sub-word model separately from a complete recognizer, to isolate its effects from those of the observation and language models. It is also sometimes necessary to do so, when testing newer, more speculative approaches for which various engineering details have not yet been addressed. One such measure is performance on the task of lexical access (also sometimes referred to as "pronunciation recognition" [81]), consisting of predicting a word given a human-labeled phonetic (or any sub-word) transcription. Other measures include phonetic error rate of predicted pronunciations [10] and perplexity of surface sub-word units given the canonical units [30], [60]. These measures are not necessarily indicative of eventual performance in a complete speech recognition system, but they help to analyze the effects of different modeling choices. Some measures, such as phonetic error rate and perplexity, are difficult to compare across models that use different types of units. Here we present a sample of results on lexical access for a subset of the phonetically transcribed portion of Switchboard [13]. Table III shows the performance of a few basic baselines, a phone-based model using context-dependent decision trees (an implementation by Jyothi et al. [60] of a model similar to that of Riley et al. [30]), and several articulatory and discriminative models. The top half of the table shows that this task is not trivial: A naïve dictionary lookup, or a lookup with rules, does very poorly (though note that a complete speech recognizer with an acoustic model would recover some of the errors made by the lexical access models). The remaining results show the potential advantages of sub-phonetic features, context modeling, and discriminative learning for sub-word modeling. As these approaches have not been tested in complete speech recognizers (except for highly constrained variants, e.g. [59]), their results must be considered suggestive at this point.

Model	Error rate (%)
Baseform lookup [50]	59.3
Knowledge-based rules [50]	56.4
Baseforms + Levenshtein distance [79]	41.8
Context-independent articulatory DBN [50]	39.0
Context-dependent phone model [60]	32.1
Context-dependent articulatory DBN [60]	29.1
CRF + phonetic/articulatory feature functions [79]	21.5
Large-margin + phonetic/articulatory feature f'ns [79]	14.8

TABLE III
LEXICAL ACCESS ERROR RATES (PERCENTAGES OF INCORRECTLY CLASSIFIED WORDS) ON A PHONETICALLY TRANSCRIBED SUBSET OF THE SWITCHBOARD DATABASE.

V. DISCUSSION

The challenges of sub-word modeling are some of the factors that have kept speech recognition from progressing beyond restricted applications and beyond high-resource settings and languages. We have motivated the need for breaking up words into sub-word units and surveyed some of the ways in which the research community has attempted to address the resulting challenges, including traditional phone-based models and less traditional models using acoustic units or sub-phonetic features. Through the unifying representation of graphical models, we have noted the commonalities and differences among the approaches. We have highlighted a few of the main existing results, showing that different types of models have benefits in certain settings. We cannot yet conclude which models are preferred in which circumstances, and certain approaches are yet to be scaled up for use in state-of-the-art systems. It is important to note that many of the approaches described here, in particular most of the work cited in Tables II and III, have not entered the mainstream; the area of sub-word modeling is still actively searching for solutions to its challenges.

Certain themes are clear, however. First, the most natural ideas of expanding phonetic dictionaries, heavily studied in the late 1990s and early 2000s, are surprisingly difficult to turn into successful sub-word models. One reason is the continuous nature of pronunciation

variation. The alternative of modeling all variation at the acoustic level achieves similar, but not improved, results to phonetic dictionary expansion. The "intermediate" approaches of sub-phonetic feature models have the potential to both cover the continuum of pronunciation variation and be more robust to low-resource settings, but have yet to be tested in large-scale recognition. Modeling *context* is important—whether it is phonetic context in phone-based models [30], word-level context that changes the prior distribution of pronunciations [10], [2], [7], or articulatory context in sub-phonetic models [60]. Finally, conditional or discriminative modeling has received relatively little attention in sub-word modeling research, but can potentially improve performance significantly [35], [65], [79].

The field is starting to benefit from combining some of the ideas discussed here, in particular through much tighter coupling between sub-word modeling, observation modeling, and machine learning techniques. New work on discriminative sequence models is making it possible to incorporate much richer structure than has been possible before [65], [79], [82], [63], [64].

We have not explored all issues in sub-word modeling in detail. In particular, the interactions between sub-word modeling, observation modeling, and the choice of acoustic observations deserve more study. For example, phonetic dictionary expansion may affect different systems differently (e.g., possibly achieving greater improvements in a segment-based recognizer [33] than in HMM-based recognizers [30], [10]), but to our knowledge there have been no direct comparisons on identical tasks and data sets. We have also only briefly touched on automatic sub-word unit learning and the related task of automatic dictionary learning [39], [40], [47].

In some domains there is now an explosion of data, making it possible to learn very rich models with large context. At the same time, there is great interest in multilingual and low-resource domains, where data is scarce and parsimonious models are particularly appealing.

One of the crucial aspects of sub-word modeling, which differentiates it from other aspects of speech recognition, is that it is modeling something that is never observed: There is no way to obtain absolute ground-truth sub-word unit labels, and we do not know precisely what these units should be. However, as we have discussed here, except in rare cases (e.g., very small vocabularies), it is necessary to break up words into sub-word units and confront the resulting challenges.

REFERENCES

[1] B. H. Repp, "On levels of description in speech research," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1462–1464, 1981.

[2] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1999.

[3] J. Odell, *The Use Of Context In Large Vocabulary Speech Recognition*, Ph.D. thesis, University of Cambridge, March 1995.

[4] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[5] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of speaking style on LVCSR performance," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996.

[6] T. Shinozaki, M. Ostendorf, and L. Atlas, "Characteristics of speaking style and implications for speech recognition," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1500–1510, 2009.

[7] M. Adda-Decker and L. Lamel, "Pronunciation variants across system configuration, language and speaking style," *Speech Communication*, vol. 29, no. 2-4, pp. 83–98, 1999.

[8] M. Ostendorf, E. Shriberg, and A. Stolcke, "Human Language Technology: Opportunities and Challenges," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[9] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[10] J. E. Fosler-Lussier, *Dynamic Pronunciation Models for Automatic Speech Recognition*, Ph.D. thesis, U. C. Berkeley, 1999.

[11] H. Soltau, F. Metze, and A. Waibel, "Compensating for Hyperarticulation by Modeling Articulatory Properties," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002.

[12] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 28, no. 1, pp. 96–103, 1985.

[13] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996.

[14] M. Adda-Decker, P. B. de Mareüil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous French," *Speech Communication*, vol. 46, no. 2, pp. 119–139, 2005.

[15] D. H. Klatt, "Review of the ARPA speech understanding project," *The Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1345–1366, 1977.

[16] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 82–87, 1975.

[17] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358–366, 2001.

[18] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Computer Speech and Language*, vol. 18, no. 4, pp. 375–395, 1999.

[19] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 89–99, 2002.

[20] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition,"

in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[21] L. Deng and J. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2702–2719, 1994.

[22] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *Speech Communication*, vol. 41, no. 2, pp. 511–529, 2003.

[23] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," in *Proc. Eurospeech*, 2003.

[24] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[25] F. Metze and A. Waibel, "A Flexible Stream Architecture for ASR using Articulatory Features," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002.

[26] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3–4, pp. 303–319, 2002.

[27] H. Bourlard, S. Furui, N. Morgan, and H. Strik, "Special issue on modeling pronunciation variation for automatic speech recognition," *Speech Communication*, vol. 29, no. 2-4, 1999.

[28] ISCA, *Proceedings of the International Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, Estes Park, Colorado, 2002.

[29] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld, "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996, Supplementary Paper.

[30] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, no. 2-4, pp. 209–224, 1999.

[31] T. Holter and T. Svendsen, "Maximum likelihood modelling of pronunciation variation," *Speech Communication*, vol. 29, no. 2–4, pp. 177–191, 1999.

[32] F. Korkmazskiy and B.-H. Juang, "Discriminative training of the pronunciation networks," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1997.

[33] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Speech Communication*, vol. 46, no. 2, pp. 189–203, 2005.

[34] M. Wester, "Pronunciation modeling for ASR - knowledge-based and data-derived methods," *Computer Speech & Language*, vol. 17, pp. 69–85, 2003.

[35] O. Vinyals, L. Deng, D. Yu, and A. Acero, "Discriminative pronunciation learning using phonetic decoder and minimum-classification-error criterion," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[36] P. Karanasou, F. Yvon, and L. Lamel, "Measuring the confusability of pronunciations in speech recognition," in *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, 2011.

[37] E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo, "On the road to improved lexical confusability metrics," In *Proc. ITRW PMLA* [28].

[38] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1996.

[39] B. Hutchinson and J. Droppo, "Learning non-parametric models of pronunciation," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[40] I. Badr, I. McGraw, and J. Glass, "Pronunciation learning from continuous speech," in *Proc. Interspeech*, 2011.

[41] H. Soltau, H. Yu, F. Metze, C. Fügen, Q. Jin, and S.-C. Jou, "The ISL evaluation system for RT-03S CTS," in *Proc. RT-03S Workshop*, 2003.

[42] M. Saraçlar and S. Khudanpur, "Pronunciation change in conversational speech and its implications for automatic speech recognition," *Computer Speech and Language*, vol. 18, no. 4, pp. 375–395, 2004.

[43] T. Svendsen, K. K. Paliwal, E. Harborg, and P. O. Husøy, "An improved sub-word based speech recognizer," in *ICASSP*, 1989.

[44] T. Hain, "Implicit pronunciation modeling," In *Proc. ITRW PMLA* [28].

[45] M. Saraçlar, H. Nock, and S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, pp. 137–160, 2000.

[46] B. Varadarajan and S. Khudanpur, "Automatically learning speaker-independent acoustic subword units," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2008.

[47] C.-Y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proc. Association for Computational Linguistics (ACL)*, 2012.

[48] T. Hain and P. Woodland, "Dynamic HMM selection for continuous speech recognition," in *Proc. Eurospeech*, 1999.

[49] T. Hain, "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech Communication*, vol. 46, no. 2, pp. 171–188, 2005.

[50] K. Livescu, *Feature-based Pronunciation Modeling for Automatic Speech Recognition*, Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, September 2005.

[51] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proc. Eurospeech*, 2003.

[52] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Sydney; Australia, 1998, ISCA.

[53] S. King, T. Stephenson, S. Isard, P. Taylor, and A. Strachan, "Speech recognition via phonetically-featured syllables," in *Proc. Workshop on Phonetics and Phonology in ASR "Phonus 5"*, 2000.

[54] J. Morris and E. Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *IEEE Transactions On Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 617–628, 2008.

[55] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," *Phonetica*, vol. 49, no. 3-4, 1992.

[56] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 33, no. 2–3, pp. 93–111, 1997.

[57] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Communication*, vol. 24, no. 4, pp. 299–323, 1998.

[58] K. Livescu and J. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2004.

[59] K. Livescu, O. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman,

S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[60] P. Jyothi, K. Livescu, and E. Fosler-Lussier, "Lexical access experiments with context-dependent articulatory feature-based models," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[61] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Gesture-based dynamic Bayesian network for noise robust speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[62] H.-K. J. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, May 2006.

[63] S.-X. Zhang, A. Ragni, and M. J. F. Gales, "Structured log linear models for noise robust speech recognition," *IEEE Signal Processing Letters*, vol. 17, no. 11, pp. 945–948, 2010.

[64] G. Heigold, H. Ney, P. Lehnen, and T. Gass, "Equivalence of generative and log-linear models," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 19, no. 5, pp. 1138–1148, 2011.

[65] G. Zweig, P. Nguyen, D. V. Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. Hermansky, D. Karakos, A. Jansen, S. Thomas, S. G.S.V.S., S. Bowman, and J. Kao, "Speech recognition with segmental conditional random fields: A summary of the JHU CLSP summer workshop.," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[66] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in *Proc. Interspeech*, 2005.

[67] Y.-H. Sung and D. Jurafsky, "Hidden conditional random fields for phone recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009.

[68] G. Zweig, *Speech recognition using dynamic Bayesian networks*, Ph.D. thesis, U. C. Berkeley, 1998.

[69] J. Bilmes, *Natural Statistical Models for Automatic Speech Recognition*, PhD dissertation, U. C. Berkeley, Berkeley, CA, 1999.

[70] J. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.

[71] S. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, UK, 1996.

[72] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, MA, 2009.

[73] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning (ICML)*, 2001.

[74] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[75] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[76] A. M. Carvalho, R. Roos, A. L. Oliveira, and P. Myllymäki, "Discriminative learning of Bayesian networks via factorized conditional log-likelihood," *Journal of Machine Learning Research*, vol. 12, pp. 2181–2210, 2011.

[77] O. Çetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An articulatory feature-based tandem approach and factored observation modeling," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[78] R. Prabhavalkar, E. Fosler-Lussier, and K. Livescu, "A factored conditional random field model for articulatory feature forced transcription," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[79] H. Tang, J. Keshet, and K. Livescu, "Discriminative pronunciation modeling: A large-margin, feature-rich approach," in *Proc. Association for Computational Linguistics (ACL)*, 2012.

[80] M. Bacchiani and M. Ostendorf, "Using automatically-derived acoustic sub-word units in large vocabulary speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1998.

[81] K. Filali and J. Bilmes, "A dynamic Bayesian framework to model context and memory in edit distance learning: An application to pronunciation classification," in *Proc. Association for Computational Linguistics (ACL)*, 2005.

[82] J. Keshet and S. Bengio (eds.), *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, John Wiley and Sons, 2009.

Karen Livescu received the A.B. degree in Physics from Princeton University in 1996 and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Science (EECS) from the Massachusetts Institute of Technology (MIT) in 1999 and 2005, respectively. From 2005 to 2007 she was a Clare Boothe Luce Postdoctoral Lecturer in the EECS department at MIT. Since 2008 she has been with the Toyota Technological Institute at Chicago (TTIC), where she is now Assistant Professor. She is a member of the IEEE Speech and Language Technical Committee and a subject editor for *Speech Communication* journal.

Eric Fosler-Lussier received the B.A.S in Computer and Cognitive Studies and the B.A. in Linguistics from the University of Pennsylvania in 1993. He received the Ph.D. degree from the University of California, Berkeley in 1999; his Ph.D. research was conducted at the International Computer Science Institute, where he was also a Postdoctoral Researcher. From 2002 to 2002, he was a Member of Technical Staff in the Multimedia Communications Lab at Bell Labs, Lucent Technologies; subsequently he was a Visiting Scientist in the Department of Electrical Engineering, Columbia University. Since 2003 he has been with the Department of Computer Science and Engineering, The Ohio State University, with a courtesy appointment in the Department of Linguistics, where he is an Associate Professor and directs the Speech and Language Technologies (SLaTe) Laboratory. He is currently serving his second term on the IEEE Speech and Language Technical Committee, and is a recipient of the 2010 IEEE Signal Processing Society Best Paper Award.

Florian Metze received his Diploma degree in Theoretical Physics from Ludwig-Maximilians-Universität München in 1998, and a Ph.D. in Computer Science from Universität Karlsruhe (TH) in 2005. He worked as a Postdoctoral Researcher at Deutsche Telekom Laboratories (T-Labs) in Berlin, Germany, from 2006 to 2008. In 2009, he moved to Carnegie Mellon University's Language Technologies Institute as an Assistant Research Professor. He is a member of the IEEE Speech and Language Technical Committee, with research interests in acoustic modeling, meta-data annotation and multi-media processing.