Incorporating linguistic theories of pronunciation variation into speech – recognition models

Mari Ostendorf

Phil. Trans. R. Soc. Lond. A 2000 **358**, 1325-1338
doi: 10.1098/rsta.2000.0589

Incorporating linguistic theories of pronunciation variation into speech-recognition models

By Mari Ostendorf

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

This paper describes the use of distinctive linguistic features to represent acoustic variability of words for speech recognition. Focusing on conventional hidden Markov model technology, we review implicit use of linguistic features as questions in decision-tree design for both coarticulation and pronunciation modelling and describe possibilities for more explicit use. The importance of conditioning on (hierarchical) syllable and prosodic structure is discussed, and the problem of modelling relative timing of feature-dependent acoustic cues is raised as a key limitation of current models.

Keywords: acoustic modelling; pronunciation modelling; phonetic variation

1. Introduction

It has often been noted that automatic speech-recognition performance is much worse on spontaneous speech than on carefully articulated speech. For the best systems reporting results on the 1999 DARPA Broadcast News benchmark tests, word error rates on the spontaneous speech portion of the test set (14–16%) were nearly double those on the baseline condition comprised mainly of news announcer recordings (8–9%) (Pallett *et al*. 1999). Those sites that also participated in a workshop on conversational speech recognition a few months later reported word error rates of *ca.* 40%. Pronunciation variability has frequently been cited as a key reason for the poor performance, and McAllister *et al*. (1998) provide evidence to support this hypothesis using simulated-data experiments. Anecdotal examples of pronunciation variability abound. For example, in a 4 h phonetically transcribed subset of the Switchboard corpus, we found over 30 different pronunciations of 'and', from 'æ n d' (canonical) to 'ɛ n' (most frequent) to a nasal flap, with at least 10 different vowels observed and frequent final consonant deletion/reduction.

Not surprisingly, there have been a large number of research efforts devoted to pronunciation modelling in the last few years, including techniques that use automatic learning, hand-written phonological rules and various combinations of the two. Unfortunately, the gains from phone-based pronunciation modelling techniques have been disappointing, e.g. reducing word error rates from 40.9% to 38.5% on conversational speech (Riley *et al*. 1999). This gain represents a statistically significant improvement on a difficult task, but not the factor-of-five reduction predicted in McAllister *et al*. (1998). Of course, the factor of five is optimistic because of the match between modelling assumptions in the recognition and simulation of data, but most researchers still share the intuition that there is more to be gained from pronunciation modelling.

In automatic speech recognition, pronunciation variation is typically modelled as insertion, deletion or substitution of a *phone segment*, where the phone inventory includes approximately 40–50 basic consonant and vowel sounds like 's', 'm', 'o' and 'I' (including multiple phones for some phonemes). In contrast, phonological variation is frequently described in linguistics in terms of simple feature changes, where a *feature* characterizes categorical contrasts between speech sounds, such as 'voiced', which distinguishes 'b' from 'p', 'z' from 's', etc. and the feature 'nasal' which is associated with the group of phones 'm', 'n' and 'ŋ'. (Note that the term 'feature' has been used to mean a variety of things in the speech-processing literature, including continuous-valued articulatory parameters, acoustic correlates of distinctive features, and the acoustic measurements computed as a first stage of recognition, all of which differ from the symbolic usage intended here.) A vector of feature values can be thought of as a particular encoding of a phoneme index, so a change in one feature corresponds to a change in the phoneme. Since the 'code' was designed to cover different languages of the world, there are possible feature combinations that do not correspond to a phoneme in English.

While current recognition training techniques already use linguistic features implicitly in the definition of phone classes, there are practical reasons why explicit use of features may give different results. A goal of this paper is to overview current approaches and show how linguistic knowledge can be used to better advantage within conventional hidden Markov model (HMM) recognition technology. Our view is that, because linguistic theory of phonetic variation is far from complete, particularly in accounting for individual speaker variation, deterministic phonological rules cannot replace statistical models or even deterministically define their structure. This is especially true for conversational speech, since the controlled studies on carefully read laboratory speech do not always translate directly to the phenomena observed in casual spontaneous speech. Instead, linguistic knowledge should be incorporated via automatic training. HMMs represent a first step.

This paper argues that phonemes are too coarse a unit for representing acoustic variation in speech for two reasons. First, a good model of phonetic variation should depend on both phonetic context and on higher-level syllable and prosodic structure. With this increase in the dimensions of context conditioning, the phoneme space may be too large for robust parameter estimation. Second, the use of phonemes limits the model of timing to sequential state durations, whereas a representation of relative state timing is critically needed.

The remainder of the paper is organized as follows. In § 2, approaches for modelling coarticulation and pronunciation variation in a phone-based HMM system are described, followed by a discussion of acoustic variation in terms of linguistic features in § 3. Next, § 4 covers recent work on incorporating linguistic structure above the level of the phone, both implicitly and explicitly. The issue of relative timing of feature realization is discussed in § 5, with concluding remarks in § 6.

2. Phone-based acoustic and pronunciation modelling

In the standard statistical approach to speech recognition, the recognition problem is posed as one of choosing the word sequence that maximizes the likelihood

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w})p(\mathbf{w}),$$

where $\mathbf{x} = x_1, \ldots, x_T$ is a T-length sequence of acoustic observations (e.g. cepstral parameters) and $\mathbf{w} = w_1, \ldots, w_n$ is a hypothesized word sequence of length n. The probability function $p(\mathbf{x} \mid \mathbf{w})$ is often referred to as the acoustic model, and $p(\mathbf{w})$ is referred to as the language model. The acoustic model typically includes three main components. First, a base lexical representation, typically called a 'baseform,' is expanded into a list of pronunciations or a pronunciation network, annotated with pronunciation probabilities. Here, a 'pronunciation' is a sequence of phone symbols. Second, each phone in the list or network is mapped to a sequence of model indices depending on its phonetic context. Lastly, a probability distribution describes the likelihood of a sequence of (continuous) acoustic observations given the model index sequence. The observation model is most often a Gaussian or Gaussian-mixture distribution, as in an HMM, but it could also be a more complex segmental distribution model or a neural network. Mathematically, these components are evaluated together in computing the probabilistic evidence for a word,

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{w}) &= \sum_{\phi} p(\phi \mid \mathbf{w}) p(\mathbf{x} \mid \phi) \\ &= \sum_{\phi} p(\phi \mid \mathbf{w}) \sum_{\mathbf{s}} p(\mathbf{s} \mid \phi) p(\mathbf{x} \mid \mathbf{s}) \\ &\approx \max_{\phi, \mathbf{s}} p(\phi \mid \mathbf{w}) p(\mathbf{s} \mid \phi) p(\mathbf{x} \mid \mathbf{s}), \end{aligned}$$

where ϕ is a pronunciation (a sequence of phones: $\phi_1, \ldots, \phi_m$), $\mathbf{s} = s_1, \ldots, s_T$ is an HMM state sequence, and the approximation in the last step is made to simplify the recognition search process. Thus, there are three component models: the pronunciation model $p(\phi \mid \mathbf{w})$; the model of sub-phonetic temporal characteristics $p(\mathbf{s} \mid \phi)$; and the observation model $p(\mathbf{x} \mid \mathbf{s})$. The first component is designed to capture pronunciation differences at the phone level, such as 'æ n d' versus 'æ n' versus 'ə n' for 'and', while the second component models coarticulation effects such as formant trajectory changes at vowel onsets and offsets. The existence of these two components demonstrates that phonetic variation takes a wide range of forms. Since linguistic features provide a good framework for understanding both extremes, and since the two components can be merged, this paper will cover both.

The aspect of modern speech recognizers with the longest history of using linguistic insights is the second component: modelling of coarticulation via context-dependent distributions, as in triphones where the phone model is conditioned on the left and right phonetic context. Distribution clustering is used to estimate models for triphones, because there are too many to estimate reliably. Clustering is typically at the level of phone states, with 3–5 sequential states per triphone to capture temporal variability. The most popular approach to distribution clustering uses decision trees with linguistically motivated questions (Young *et al.* 1994). In other words, hand-specified phone classes (e.g. grouped by manner and/or place of articulation) define a set of binary questions, and the automatic decision-tree design algorithm chooses to split groups of context-dependent models according to the question that results in the greatest increase in likelihood.† In other words, the state s in $p(x \mid s)$ is indexed

† Note that this use of decision trees is slightly different than the standard use, described by Breiman *et al.* (1984) and in the pronunciation modelling discussion below, in that the objective is maximum likelihood of data from a continuous-valued vector variable, rather than minimum entropy of the empirical distribution for a discrete (categorical) variable.

by a decision-tree leaf node, $s = \mathcal{T}(\phi)$. Alternative data-driven clustering algorithms have been proposed, but an advantage of using linguistic classes is that the models typically generalize well to contexts that are unseen in training data. In other words, if a particular vowel is seen in the training data followed by both 'n' and 'm' but not 'ŋ', then the effect of nasalization can be learned for all by defining a nasal class. Most clustering algorithms assume a fixed state topology for all triphones, but improved temporal modelling can be achieved by allowing splitting as a function of temporal position as well as a function of neighbouring phonetic context (Ostendorf & Singer 1997). Clustering context-dependent models is very effective for modelling certain types of acoustic variation, but it cannot handle phenomena like apparent segment deletion, since the assumption is that every context-dependent phone is realized with some minimum duration (*ca.* 30 ms). Substitution can be handled by using mixture distributions in the observation model, but this is a weak model of pronunciation variation that allows implausible pronunciations (e.g. switching phones midway through the segment).

Explicit pronunciation modelling, in the sense of predicting alternate phone sequences for a word, has become an active area of research as systems have matured and been applied to spontaneous speech. Phonological knowledge is incorporated in a statistical model in two main ways. One strategy involves training probabilities of a set of hand-written context-dependent phonological rules (Cohen 1989; Tajchman *et al.* 1995). A variation of this approach involves learning the context conditioning for rule probabilities using a decision tree (Finke & Waibel 1997). In these cases, the probability of a pronunciation is determined by the product of the probabilities of the rules used to derive it. An alternative is to use decision trees to predict realized phone identities given the baseform phone sequence (Riley *et al.* 1999), in which case the word pronunciation probability is given by the product of the predicted phone probabilities (from the tree leaf nodes),

$$p(\phi \mid w) = \prod_j p(\phi_j \mid T(\phi_{j-1}, w)), \tag{2.1}$$

where $T(\cdot)$ is the decision tree and w includes the base pronunciation and lexical stress pattern of the word. The methods share the technique of building an initial set of pronunciations (based on human knowledge or hand-transcribed data), using forced alignment to determine which pronunciation is used for each instance of a word in a large set of training data, and then training a new pronunciation model based on these phone labels. A problem with this approach, nicely illustrated by Saraclar *et al.* (1999), is that improving the phone transcription via the forced alignment step may lead to better phone-recognition models but possibly poorer word recognition. The study by Riley *et al.* (1999) may explain this in part: the assumption of conditional independence used in multiplying phone-realization probabilities is an oversimplification that leads to poor word-level pronunciation probabilities.

In summary, linguistic knowledge is already widely (and successfully) used in speech-recognition systems, though the use of linguistic features *per se* is mostly implicit in the definition of questions for decision-tree design. Next, we look at phonological variation from a linguistic perspective to see if there might be more to be gained from explicit use of distinctive features in speech recognition.

3. Distinctive features, phonological variation and prosody

In automatic speech recognition, the basic building blocks are phonemes (or phones), which are divided into sub-phonetic regions that are sequential in time. In linguistics, phonological features are typically viewed as the fundamental building blocks of speech (Halle 1992), and phonemes are specified (or coded) in terms of features with little or no representation of time. For the most part, distinctive features are related to the manner in which a speech sound is produced (the degree of constriction in the vocal tract), the particular articulator that is used (glottis, soft palate, lips and tongue blade, body and root) and/or place of constriction, and how an articulator is used to produce the sound.† Different feature systems have been proposed, including binary and multi-valued features, as discussed in Clark & Yallop (1995). Examples of binary features are *nasal, voiced, continuant, round*, etc. An example of a multi-valued feature might be place of articulation, taking on values *velar, dental, labial*, etc. Some binary features are values in a multi-valued system, e.g. *nasal* and *continuant* are possible values of the feature 'manner'.

Distinctive features are associated with acoustic correlates, though not all of these are well understood. The correlates may also depend on combinations of features. For example, the feature *voiced* is generally associated with periodicity in the time signal, but one cue to a voiced stop consonant is a shorter time from the start of the burst to the onset of voicing than for the unvoiced counterpart.

Pronunciation variations are sometimes expressed in terms of context-dependent rules describing changes in the feature values or in feature association with segments. Features may change values, as in a change from + to − when a vowel or final consonant is devoiced in the context of a subsequent voiceless consonant, and when a tense vowel 'i' becomes a lax 'I'; or a change in the place of articulation, as when 'n' becomes 'm' when followed by a labial stop (as in 'can be' or 'grampa'). In a feature system that uses the notion of unspecified features as a third 'value' of an otherwise binary feature (Lahiri 1999), vowel reduction can be thought of as changing a feature value to be unspecified. Feature changes can lead to situations where phone segments appear to be deleted when there is still evidence for these segments in the realization of neighbouring segments, as in a nasalized 'æ' in a reduced form of 'ca[n]'t' or the single nasal–dental segment sometimes produced for the two consonants in 'in the'. The features that define a phoneme do not always map to acoustic cues that form synchronous parallel time functions, which can explain cases where segments appear to be inserted, as in an epenthetic stop in 'warm(p)th' due to asynchronous alignment of the nasal and continuant feature cues. These sorts of feature changes, sometimes referred to as 'feature spreading',‡ can raise significant problems for phone-level transcription of spontaneous speech (Fosler-Lussier *et al.* 1999).

The fact that certain sets of features tend to spread or reassociate as a group has been used to argue for a hierarchical organization of features (Clements 1985). Different hierarchies have been proposed; figure 1 (from Keyser & Stevens (1994)) illustrates a feature geometry motivated by the structure of the vocal tract. As we

† Although distinctive features are closely related to articulation, they are themselves abstract dimensions. In particular, they are not articulatory parameters in the sense of the 'features' used in work by Deng and co-workers (see, for example, Erler & Deng 1993; Deng & Wu 1996), which are inherently continuous but quantized for purposes of defining discrete HMM states.

‡ The term 'feature spreading' has theoretical connotations that we would like to avoid here, but its non-technical interpretation is useful for visualizing consequences for HMM state definitions.

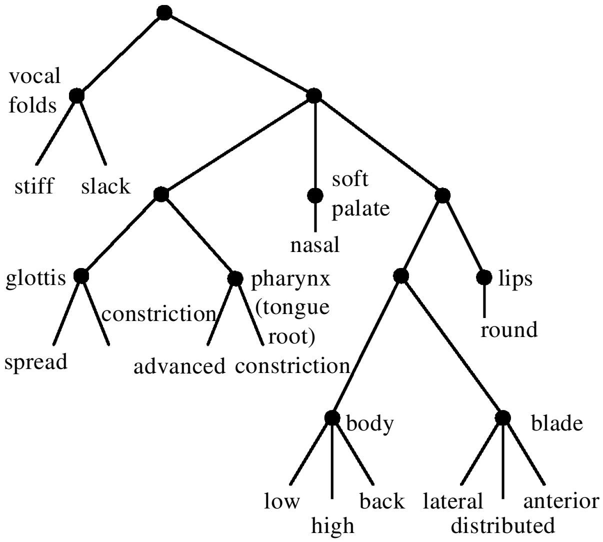

Figure 1. A feature hierarchy motivated by vocal tract structure, with features labelled at leaves and articulators labelled at internal nodes (reproduced from Keyser & Stevens (1994)).

will point out later, the existence of a hierarchy has important implications for speech recognition, because of the possibility for parsimonious representation of statistical dependence. While the hierarchy suggests some degree of independence between different 'mini-tracts' of the vocal tract, there are interactions between some features that *enhance* certain phonetic contrasts (Stevens & Keyser 1989). Such interactions imply that acoustic observation models should be conditioned on sets of features and not only on individual features.

Pronunciation variation (and, therefore, the probability of feature changes) appears to be very much dependent on syllable structure. Based on an analysis of hand-labelled phonetic transcriptions of the Switchboard corpus, Greenberg (1998) observes that syllable onsets are most often canonical and codas are most frequently changed or deleted. In a comparison of the conversational Switchboard data to the read speech in the TIMIT corpus, the biggest difference is in the variability of the coda consonants (Fosler-Lussier *et al.* 1999).

In addition, there appears to be evidence that higher-level structure also plays a role in the likelihood of feature changes, including word frequency, syntax and/or prosodic factors. Fosler-Lussier *et al.* (1999) show an interaction between speaking rate and word frequency in predicting how much a word pronunciation will deviate from a dictionary baseform. Syntax appears to be a factor as well: it would sound strange to have 'did you' spoken as '**d ɪ ǰ ə**' at a major syntactic clause boundary (as in 'If I did, you...'). However, such phenomena may be more directly described in terms of prosodic structure (Shattuck-Hufnagel & Turk 1996), which is related to (but not identical to) syntactic structure. Cross-word-boundary phonological changes, as in the '**d ɪ ǰ ə**' example, typically do not occur at major prosodic phrase boundaries, and other insertion-like effects do occur at prosodic boundaries. Dilley *et al.* (1996) found that glottalization was more likely at vowel–initial word onsets when those words were pitch-accented and/or the first word of a prosodic phrase. The frequency of glottalization increased with increased saliency of the location, such that glottalization was quite likely (above 90% for the female subjects) if a word was both accented and phrase-initial. Fougeron & Keating (1997) measured increased tongue contact with the palate during 'n' for initial consonants of prosodic constituents in reiterant speech. Such articulatory strengthening presumably has an acoustic consequence, as illustrated by the difference in consonant bursts as a function of syllable and word position. There may be an effect of enhanced phonetic realization via inserted features at particularly salient regions of the speech signal: hyperspeech in the 'hyper and hypo' (H&H) theory (Lindblom 1990). In the Switchboard corpus, there are at least anecdotal examples, e.g. an off-glide of 'æ' is enhanced in an emphasized pronunciation of 'and', resulting in 'æ **ɛ** n d' (using a phonetic alphabet). We conjecture that conditioning feature changes on a prosodic hierarchy, starting from the level of the syllable, will be needed to better explain the pronunciation variability in speech.

4. Modelling higher-level structure with HMMs

When all the observed pronunciations of a word are allowed in speech-recognition decoding, performance degrades due to the increased confusability between words, e.g. allowing 'æ n' as a pronunciation for 'and' increases the possibility of confusing 'and' and 'an'. For this reason, the dependence of pronunciation variability on higher-level linguistic structure is of great importance to speech-recognition systems—it provides a means of dynamically varying pronunciation probabilities. Researchers have begun exploring methods for introducing higher-level structure within the context of the standard statistical (i.e. HMM) recognition paradigm, taking advantage of multi-pass search architectures to condition on hypothesized word context. This section will describe the two main developments, corresponding to the distribution clustering and pronunciation modelling components described earlier. In both cases, linguistic features are again used only implicitly, which we will argue may be limiting the success of the extensions.

In order to incorporate syllable structure directly into design of the acoustic model index sequence, an extension of the standard HMM context-dependent model clustering framework was developed, referred to as *tagged clustering*. Tagged clustering incorporates symbolic descriptions of a base phoneme that reflect higher-level context, making it possible to capture phenomena such as a tendency to reduce unstressed vowels and to more strongly release a stop consonant in word onset position. Each phone in a dictionary is tagged according to factors like lexical stress, syllable position, word position, etc. Then, tri-tag models are trained and clustered, just as for triphone models, except that the decision tree must choose between questions that are motivated by these tags as well as those defined in terms of phonetic context. The idea of tagged clustering was first introduced in speech synthesis by Donovan (1996), who found that lexical stress was among the most important questions in the sense of being asked early in the tree. The importance of stress has also been observed in recognition experiments by others (Ostendorf *et al.* 1997; Paul 1997). Word position (beginning, middle, end) has been found to be important in experiments by several researchers. The usefulness of syllable position is unclear; our unpublished experiments contradict the negative results reported by Paul (1997), perhaps because of our representation of multiple types of syllable onsets. A limitation of tagged clustering is that coding phones causes a huge increase in the number of elementary context-dependent models, which leads to large memory requirements and increased complexity of training because of the increase in possible data divisions. As a result, only simple tag sets have been explored in large vocabulary systems using cross-word context. Work in progress on multi-stage clustering may address this problem by using different subsets of features in different stages of tree design.

The same higher-level tags can be used more easily in *decision-tree pronunciation modelling*. Already, syllable structure and stress have proved to be useful (Weintraub *et al*. 1996; Riley *et al*. 1999), but the problem of independence assumptions raised in § 2 remains. An interesting solution to this problem is proposed by Fosler-Lussier *et al*. (1999); they predict syllable-level pronunciations using decision trees, which gives a reduction of *ca*. 10% WER of the spontaneous speech portion of the DARPA Broadcast News task. They allow questions on syllable structure, as well as hypothesized local word context, speaking rate, etc. Finke & Waibel (1997) have investigated conditioning on similar factors in decision trees used to predict rule probabilities, obtaining significant gains in both phone prediction and word-recognition performance over using local phonetic context alone. Yet another approach is to condition pronunciation probabilities, either word-level or decision-tree-based, on a discrete hidden speaking mode variable predicted from acoustic cues and the hypothesized word sequence (Ostendorf *et al*. 1997). The hidden mode can be thought of as a mapping of high-level conditioning factors to a small space via unsupervised clustering. While prosodic structure has not been used directly in any of this work, it has been used indirectly via acoustic cues (such as presence of a pause), which may indicate a prosodic phrase boundary.

In the above extensions, linguistic feature theory is not used explicitly; features are implicit in the definition of phonetic classes for decision-tree question learning. Given infinite training data, one might argue that there is no difference between implicit and explicit use of linguistic features. After all, features simply provide a particular encoding of phonemes. However, the reality is that training data are limited; experiments in Riley *et al*. (1998) show that recognition performance actually degrades if pronunciation models are trained on a small subset of hand-labelled data. The problem is exacerbated by conditioning on higher-level factors, which necessarily occur less frequently than the triphones used in current context-dependent models. Linguistic features provide a lower-dimensional representation for pronunciation prediction that can be more efficiently trained, i.e. estimating the probability of a binary feature change (1 parameter) requires less data than estimating the probabilities of 40–50 phones. Another motivation for explicit use of symbolic linguistic features in HMMs is the potential for incorporating (and optimizing) signal processing to extract feature-motivated correlates (Bitar & Espy-Wilson 1996; Kirchhoff 1996, 1998; King *et al*. 1998), which are potentially more robust than standard cepstral features and are likely to generalize better across languages.

A key question is: to what extent can one assume independence of features or subsets of features? The simplest approach is to replace $p(\phi_j \mid \phi_{j-1}, w)$ in equation (2.1) with a product of feature terms,

$$p(\phi_j \mid \phi_{j-1}, w) = \prod_{k=1}^{d} p(f_{j,k} \mid f_{j-1,k}, w), \tag{4.1}$$

where $f_{j,k}$ is the kth element of the feature vector at time j, and w is coded in terms of features rather than phones. Unfortunately, this solution ignores the interdependence of feature changes and further exacerbates the problems of conditional independence of phones. The hierarchical description of features may be useful here for specifying a Markov-like dependence tree that allows conditioning feature changes on feature values higher in the tree, as in

$$p(\phi_j \mid \phi_{j-1}, w) = \prod_{k=1}^{d} p(f_{j,k} \mid f_{j,\pi(k)}, \mathbf{f}_{j-1,h(k)}, w), \tag{4.2}$$

where $\pi(k)$ is the 'parent' of k in the tree hierarchy and $\mathbf{f}_{j-1,h(k)}$ is the sub-vector of features that are important for predicting the kth element of the feature vector.

In the above equations, we omitted the decision-tree dependence in the conditioning space for notational simplicity. Decision trees $T_k[\cdot]$—one for each of the d elements of the feature vector—are incorporated as in $p(f_{j,k} \mid T_k[f_{j,\pi(k)}, \mathbf{f}_{j-1,h(k)}, w])$. This dependence of variables both within a tree and across time is similar to the hidden Markov decision trees proposed by Jordan *et al*. (1996), though here we make use of two trees: decision trees (for questions about w) and feature hierarchies. The decision tree can automatically learn the appropriate sub-vector $h(k)$, and also allows use of higher-level structure. The success of such a model at predicting observed feature changes can be used to evaluate different feature hierarchies.

5. Issues of timing

A limitation of all of the above approaches is in the modelling of relative timing, since features cannot be mapped to a bank of synchronously changing acoustic cues. At issue here is not the sophistication of the segmental duration model, though HMMs are known to have weak duration models, but that a more fine-grained control of temporal variability is needed than the fixed number of states per phone-sized unit used in most systems. Recognition experiments showing improved performance from using context-dependent HMM triphone topologies support this claim, and timing studies for speech synthesis also point to the need for sub-segmental duration modelling (Van Santen 1997).

In the standard HMM framework, there have already been some efforts at modelling pronunciation variability at a finer-grained time-scale, e.g. the model index sequence. The idea here is that, rather than substituting one entire phonemic segment for another, which can influence the choice of models for three segments because of context-dependent modelling, partial segment substitution or deletion is allowed. State-level pronunciation modelling has been explored without the use of any linguistic knowledge (Eide 1999; Saraclar *et al*. 1999), showing gains over phone-level pronunciation models with a more compact representation. Neither of these approaches makes use of linguistic features, but it is easy to imagine doing so in the decision-tree prediction paradigm.

An alternative to data-driven HMM state-level pronunciation modelling is to represent asynchronous acoustic cues as resulting from asynchronous distinctive feature 'state' changes, as illustrated in figure 2 for a binary feature encoding.† Deng & Erler (1992) proposed a set of parallel linguistic feature streams, with rules for constraining feature 'spreading,' that are compiled together into what is effectively a

† Representing features using a time-dependent state is not consistent with the linguistic notion of a feature as an abstract discrete event, but it is useful for the HMM implementation.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

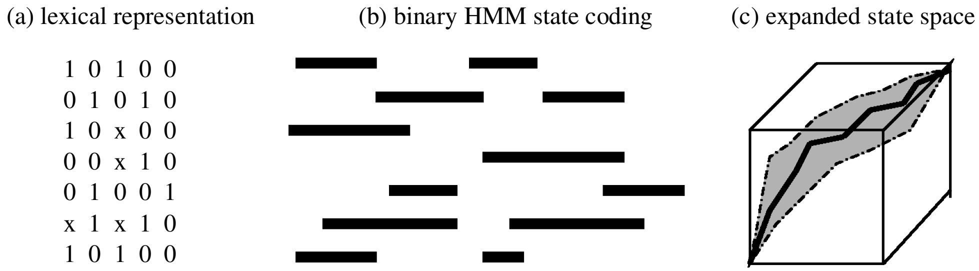

Figure 2. Conceptual illustration of a system where a binary encoding of phonemes (with unspecified features indicated by 'x') maps to parallel, asynchronous binary feature streams, which can be interpreted as a path in a d-dimensional state space. The shaded area indicates regions of constraints that might be specified given higher-level structure.

context-dependent HMM with state sharing determined by human knowledge rather than automatic clustering. A limitation of the approach is that independent training of the composite states corresponds to assuming that all feature dimensions are interdependent; there is no mechanism for training unseen states. More recent work looks at extending triphone clustering techniques to this paradigm, though with limited success (Deng & Wu 1996). The training problem can be addressed by treating the different features and their associated acoustic parameters as independent streams, using two-level (product state space) HMM decoding with synchronization of the streams at the syllable level (Kirchhoff 1996; King *et al.* 1998). Treating the streams as independent also simplifies the problem of decoding the high-dimensional state space. In addition, the framework nicely accommodates a variety of different acoustic measures, which can lead to improved performance in high noise (0 dB) conditions (Kirchhoff 1998).

The independence assumption can lead to too much flexibility, however, as evidenced by the fact that a more traditional phone-based model outperforms the feature-based system in low-noise conditions (Kirchhoff 1998). Two main problems stand out. First, the independent decoding of the different feature streams within the syllable corresponds to the independence assumptions in equation (4.1), which is problematic because of the interdependence of feature changes. The tree-based state prediction model in equation (4.2) provides more constraints, but at the cost of higher decoding complexity. Second, the acoustic correlates of the different features are not strictly independent, as mentioned earlier with respect to 'enhancement'. Such interactions imply that acoustic observation models should be conditioned on subsets of features and not individual features. The work of Bilmes (1999) on learning model structure may provide an automatic mechanism for learning an appropriate dependence structure that also keeps the model dimensionality small.

All of these finer-grained modelling techniques ignore the higher-level conditioning factors argued for in § 3, and one might think that the need for low-level variability modelling is at odds with the call for high-level context conditioning. Yet there is growing evidence that the relative timing of gestures is related to higher-level prosodic structure. For example, the work of Beckman and co-workers (see, for example, Edwards & Beckman 1988) demonstrates that timing is influenced by prosodic prominence and phrase structure. In other words, feature 'state' changes may be asynchronous, but the relative timing is not completely unconstrained. In fact, it appears to be highly systematic with respect to higher-level structure.

By better modelling the relationship between feature spreading (or reassociation) and relative timing in different contexts, the amount of allowed pronunciation variability can be dynamically constrained. Thus, the issue of relating timing control to higher-level structure may be one of the most important problems to address in modelling phonological variation. Adjusting HMM state transitions according to equation (4.2) is one solution, but state-transition probabilities are weak relative to the high-dimensional observation models typically used. Learning context-dependent constraints on temporal warpings, as allowed in a segmental model (Ostendorf *et al*. 1997), may provide another solution.

6. Conclusions

In summary, we have reviewed how current recognition technology already makes implicit use of linguistic features in conventional HMMs. In both pronunciation modelling and context-dependent distribution clustering, linguistic knowledge is used to define allowable questions for decision-tree design, which automatically determines the importance and interdependence of these factors. However, we argue that there are greater gains to be had by using higher levels of linguistic structure in conditioning phonological variation, and by modelling variation at a sub-segment level. While this remains to be shown experimentally, we conjecture that the explicit use of distinctive features in pronunciation modelling will facilitate fine-grained modelling, but that more sophisticated models of timing are also needed.

In this paper, we have taken the position that much can be done within the confines of conventional hidden Markov modelling and its derivatives. This is an important place to start, because it offers a wealth of existing tools and knowledge to build on, and, therefore, a near guarantee of improving over the state of the art. However, we also note that there are alternative models that may match the event-driven linguistic feature view of the speech process better (e.g. Hübener & Carson-Berndsen 1994; Stevens 1995; Niyogi *et al*. 1998), though there are statistical-modelling and efficient-decoding questions still to be resolved with these frameworks. In addition, we have chosen not to use explicit articulatory features, in part because their essentially continuous nature is not so well suited to a discrete-state model, but also because the possibility of multiple articulatory configurations for certain sounds greatly complicates the model. Again, there is interesting work in this direction attempting to address these problems (Deng 1998).

One of the advantages of using linguistic knowledge in statistical modelling, in addition to the potential for improved performance and better generalization, is the possibility of actually increasing our knowledge based on the automatically learned structure of the resulting model. So far, most of what we see in the automatically learned structure is not at all surprising to linguists, e.g. that lexical stress and syllable position affect pronunciation variability. However, the fact that such structure can be learned is at least promising, given current gaps in human knowledge. Particularly for spontaneous speech, where it is difficult to design controlled experiments, our understanding of the interaction between prosodic and segmental or sub-segmental structure may be advanced by the ability to analyse large amounts of data with statistical models.

The author thanks Katrin Kirchhoff for comments on the manuscript. This research was supported by the US National Science Foundation, grant no. ISI-9618926. The views and conclusions contained in this document are those of the author and should not be interpreted as reflecting the official policies of the funding agency.

References

Bilmes, J. 1999 Natural statistical models for automatic speech recognition. PhD thesis, Department of EECS, University of California, Berkeley, USA.

Bitar, N. & Espy-Wilson, C. 1996 Knowledge-based parameters for HMM speech recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 29–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. 1984 *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks.

Clark, J. & Yallop, C. 1995 *An introduction to phonetics and phonology*. London: Blackwell.

Clements, G. 1985 The geometry of phonological features. In *Phonology Yearbook*, vol. 2, pp. 223–252.

Cohen, M. 1989 Phonological structures for speech recognition. PhD thesis, University of California, Berkeley, USA.

Deng, L. 1998 Computational models for speech production. In *Computational models of speech pattern processing* (ed. K. Ponting). Springer.

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

Deng, L. & Wu, J. 1996 Hierarchical partition of the articulatory state space for overlapping-feature based speech recognition. In *Proc. Int. Conf. Spoken Language Processing*, pp. 2266–2269.

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

Donovan, R. 1996 Trainable speech synthesis. PhD thesis, University of Cambridge, UK.

Edwards, J. & Beckman, M. 1988 Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica* **45**, 156–174.

Eide, E. 1999 Automatic modeling of pronunciation variations. In *Proc. DARPA Broadcast News Workshop*, pp. 95–98.

Erler, K. & Deng, L. 1993 Hidden Markov model representation of quantized articulatory features for speech recognition. *Comp. Speech Language* **7**, 265–282.

Finke, M. & Waibel, A. 1997 Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proc. Eur. Conf. Speech Communication and Technology*, pp. 2379–2382.

Fosler-Lussier, E., Greenberg, S. & Morgan, N. 1999 Incorporating contextual phonetics into automatic speech recognition. In *Proc. Int. Congr. Phonetic Sciences*, pp. 611–614.

Fougeron, C. & Keating, P. 1997 Demarcating prosodic groups with articulation. *J. Acoust. Soc. Am.* **97**, 3384.

Greenberg, S. 1998 Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. In *Proc. ESCA Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition*, pp. 47–56.

Halle, M. 1992 Phonological features. In *International encyclopedia of linguistics* (ed. W. Bright). Oxford University Press.

Hübener, K. & Carson-Berndsen, J. 1994 Phoneme recognition using acoustic events. In *Proc. Int. Conf. Spoken Language Processing*, pp. 1919–1922.

Jordan, M., Ghahramani, Z. & Saul, L. 1996 Hidden Markov decision trees. MIT Computational Cognitive Science technical report 9606.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Keyser, S. & Stevens, K. 1994 Feature geometry and the vocal tract. *Phonology* **11**, 207–236.

King, S., Stephenson, T., Isard, S., Taylor, P. & Strachan, A. 1998 Speech recognition via phonetically featured syllables. In *Proc. Int. Conf. Spoken Language Processing*, pp. 1031–1034.

Kirchhoff, K. 1996 Syllable-level desynchronisation of phonetic features for speech recognition. In *Proc. Int. Conf. Spoken Language Processing*, pp. 2274–2276.

Kirchhoff, K. 1998 Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proc. Int. Conf. Spoken Language Processing*, pp. 891–894.

Lahiri, A. 1999 Speech recognition with phonological features. In *Proc. Int. Congr. Phonetic Sciences*, pp. 715–718.

Lindblom, B. 1990 Explaining phonetic variation: a sketch of the H&H theory. In *Speech production and speech modelling* (ed. W. Hardcastle & A. Marchal), pp. 403–439. Dordrecht: Kluwer.

McAllister, D., Gillick, L. Scattone, F. & Newman, M. 1998 Fabricating conversational speech data with acoustic models: a program to examine model–data mismatch. In *Proc. Int. Conf. Spoken Language Processing*, pp. 1847–1850.

Niyogi, P., Mitra, P. & Sondhi, M. M. 1998 A detection framework for locating phonetic events. In *Proc. Int. Conf. Spoken Language Processing*, pp. 1067–1070.

Ostendorf, M. & Singer, H. 1997 HMM topology design using maximum likelihood successive state splitting. *Comp. Speech Language* **11**, 17–42.

Ostendorf, M. (and 12 others) 1997 Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. Available from http://www.clsp.jhu.edu/ws96/.

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

Paul, D. 1997 Extensions to phone-state decision-tree clustering: single tree and tagged clustering. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 1487–1490.

Riley, M. (and 10 others) 1999 Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Commun.* **29**, 209–224.

Saraclar, M., Nock, H. & Khudanpur, S. 1999 Pronunciation modeling by sharing Gaussian densities across phonetic models. In *Proc. Eur. Conf. Speech Communication and Technology*, pp. 515–518.

Shattuck-Hufnagel, S. & Turk, A. 1996 A prosody tutorial for investigators of auditory sentence processing. *J. Psycholing. Res.* **25**, 193–247.

Stevens, K. 1995 Applying phonetic knowledge to lexical access. In *Proc. Eur. Conf. Speech Communication and Technology*, pp. 3–11.

Stevens, K. & Keyser, S. 1989 Primary features and their enhancement in consonants. *Language* **65**, 81–106.

Tajchman, G., Fosler, E. & Jurafsky, D. 1995 Building multiple pronunciation models for novel words using exploratory computational phonology. In *Proc. Eur. Conf. Speech Communication and Technology*, pp. 2247–2250.

Van Santen, J. 1997 Segmental duration and speech timing. In *Computing prosody* (ed. Y. Sagisaka, N. Campbell & N. Higuchi). Springer.

Weintraub, M., Fosler, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraclar, M. & Wegmann, S. 1996 WS96 project report: automatic learning of word pronunciation from data. Available from http://www.clsp.jhu.edu/ws96/.

Young, S., Odell, J. & Woodland, P. 1994 Tree-based state tying for high accuracy acoustic modelling. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 307–312.

Discussion

K. I. B. Spärck Jones (*University of Cambridge, UK*). You said that in some areas, such as prosody, linguistic theory is simply lacking. What are those areas?

M. Ostendorf. One is the relationship between feature changes and prosodic structure: we know that there *are* effects, but we do not have a very good understanding of this yet. Another problem is variability between individuals. We have found that to be quite extensive and apparent, but it has not been well studied.

S. Isard (*University of Edinburgh, UK*). How would you deal with the obvious differences *between* speakers, such as the difference between speakers who have postvocalic /r/ and those who do not?

M. Ostendorf. I think that we need adaptive models to deal with such cases.

E. Janke (*IBM, UK*). Could your system be improved by improving phone-recognition accuracy?

M. Ostendorf. That is not such an interesting strategy: improving the phone accuracy does not always improve the word model. Optimizing the performance of a lower level of analysis can even detract from success in recognizing higher-level units, which is the real goal.