A Linguistic Feature Representation of the Speech Waveform

Ellen Eide* Sanjoy Mitter†

1 Introduction

Linguists define a phoneme as a shorthand notation for a set of features which describe the operations of the articulators required to produce the meaningful aspects of a speech sound. In this paper we discuss a method of representing the speech waveform in terms of the same set of distinctive linguistic features, rendering it appropriate for a linguistically-motivated method of lexical access.

Spoken words are composed of phonemes in the same manner that written words are composed of letters; as handwritten script bears the characteristics of an individual writer, acoustic realizations of phonemes bear characteristics specific to an individual speaker. For the sake of minimal effort in generation, both handwritten text and continuous speech trains are subjected to a deformation of the individual building blocks of the message in order to smoothly link components into a unified chain.

However, in both handwriting and speech, the variations to the prototypical building blocks must not be so large as to distort the inherent qualities of the units if the result is to be understandable by other individuals. This fact suggests a description of the characters in terms of a set of attributes which are preserved under allowable deformations of the generic unit.

Just as letters may be described in terms of the strokes of the pen needed to produce them or by the manifestations of these actions such as line segments and curves, phonemes may be described in terms of the actions of a speaker needed to produce them, as well as time-varying frequency spectra which result from these actions.

In this paper the second representation is viewed as a set of observations which provide information about the first; estimates of the speaker-independent actions required to produce a sound are derived from the speaker-dependent

*BBN Systems and Technologies
†Massachusetts Institute of Technology

acoustic manifestations. We construct a flexible framework whose structure mirrors phenomena which occur in the speech signal to represent the speech waveform in terms of the speaker actions used to describe abstract speech sounds. Statistical models of arbitrary complexity may be incorporated into the general framework.

Although the introduction of an intermediate level of abstraction between the physical and lexical representations of speech corresponds to a loss of information in an information theoretic sense, the goal of a representation for the purposes of recognition is not necessarily to preserve the raw information content of the signal but to concisely capture those aspects of it which reflect distances from lexical items in a form which can be readily modeled. The success of our simple Gaussian models, as demonstrated in section ??, highlights the fact that the underlying framework is capturing aspects of the signal which convey information relevant to phonemic distinctions.

In section ?? we discuss physical and abstract representations of speech and their implications for modeling and lexical access. In section ?? we describe our framework for parameterizing the speech waveform in terms of linguistic features. This representation of the speech waveform is appropriate for lexical access on the basis of features in the task of continuous speech recognition; an application of this method of lexical access, secondary classification in keyword spotting, is described in section ??.

2 Representations of Speech

Related to the method of lexical access enabled by any representation is the notion of distance between phonemes implied by it. Most automatic speech recognition systems represent lexical entries in terms of a phonemic spelling and access words in terms of sequences of phonemes. That representation, however, disregards some of the phenomena which occur in conversational speech. In particular, relaxation of requirements on the production of a particular feature may occur. The following discussion is patterned after one given by Stevens [?]. Consider the expression "did you" which, when pronounced carefully, corresponds to the phonemes [D-IH-D-Y-UW]. When pronounced casually, however, the result may correspond to the phonemes [D-IH-JH-UH]. Phonemically, a considerable change has taken place in going from the theoretical representation of the expression and the representation corresponding to the utterance produced. Table ?? provides a representation of each of the pronunciations in terms of linguistic features, as will be described in section ??. In the feature representation of the utterances, we see that the matrix entries remain largely intact in going from the first pronunciation to the second, with only the features anterior and strident changing in the collapsing of the D-Y to JH and the feature tense changing in the final vowel. The task of recovering the word sequence is more

tractable from the second representation than from the first, since in the feature representation distance reflects directly phonemic differences, while distance in the waveform space is taken as geometric distance between spectra which may be swamped with differences which are not phonemically relevant. This claim is verified by our experimental results of section ?? in which performing lexical access on the basis of features rather than phonemically improves discrimination among potential occurrences of a word of interest.

As another example, while one may feel that the phonemes "m" and "b" are close in some perceptual space, these sounds are quite different spectrally. In the feature representation, however, they differ in only one feature, so that the intuitive proximity is captured.

	D	IH	D	Y	UW	D	IH	JH	UH
VOCALIC	-	+	-	-	+	-	+	-	+
CONSONANTAL	+	-	+	-	-	+	-	+	-
HIGH	-	+	-	+	+	-	+	+	+
BACK	-	-	-	-	+	-	-	-	+
LOW	-	-	-	-	-	-	-	-	-
ANTERIOR	+	-	+	-	-	+	-	-	-
CORONAL	+	-	+	-	-	+	-	+	-
ROUND	-	-	-	-	+	-	-	-	+
TENSE	-	-	-	-	+	-	-	-	-
VOICE	+	+	+	+	+	+	+	+	+
CONTINUANT	-	+	-	+	+	-	+	-	+
NASAL	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	-	+	-
LABIAL	-	-	-	-	-	-	-	-	-

Table 1: Feature matrices for careful as well as casual pronunciations of "did you."

In section ?? we discuss a physical representation used for automatic speech recognition; in section ?? we describe an intermediate, abstract representation of speech in terms of linguistic features.

2.1 A Physical Representation

Recognition is simply a representation at a certain level of abstraction. For example, a hidden-Markov-model-based continuous speech recognition system (HMM) with a null grammar finds the most likely sequence of lexical items to represent a waveform, thereby transforming the physical representation directly to a representation at the word level. With a language model, an HMM transforms the waveform to a representation at the phrase level. As the HMM goes

directly from the physical representation to one in terms of lexical items, lexical access is necessarily performed on the basis of the physical features.

Data Reduction

We begin our discussion of HMM representations with a description of the physical features used in its models. The speech waveform is subjected to a data-reduction stage whereby an attribute vector is extracted every 10ms to describe the signal. At each frame, the waveform is assumed to be stationary over a 20ms Hamming window. The short-time Fourier transform of each windowed section of the waveform is computed and the logarithm of the squared magnitude of the spectrum is retained. The frequency axis is warped according to a mel transformation [?] to imitate the processing performed in the human auditory system, where spectral components at low frequencies are replicated with higher spectral resolution than those which occur in high frequency regions. The inverse FFT of the result is truncated to 14 cepstral coefficients to form the basic signal representation in the neighborhood of the current frame. The Hamming window is advanced 10ms and the computation is repeated to form the representation of the next frame.

The average cepstral vector is computed for each utterance, and that average is subtracted from each frame to provide the normalized representation used in modeling. In the case of a stationary channel response convolved with the input waveform, the result in the frequency domain is a multiplication of the channel spectrum with the input spectrum. The logarithmic operation in computing the cepstrum turns that multiplication into an additive operation which is canceled through subtraction. Here we are assuming that every utterance is long enough so that we have similar spectra in each utterance and long-term differences between utterances are due only to channel variations.

Additionally, first and second derivatives of the cepstral vector are estimated at frame for use as waveform attributes to be modeled.

Model Topology

Figure ?? shows the topology used to model the acoustic attributes which result from the utterance of a single phoneme. The Markov model consists of 5 states connected in a left-to-right progression with skips allowed.

At each state a probability density function is parameterized to model the variability in feature space incurred in physical attribute vectors aligning to that state; transitions between states model variability in time as exponentially distributed. Models for words are formed by concatenating the models of the constituent phonemes.

Implicit in the Markov modeling of words in terms of physical attribute vectors lies the assumption that the physical features are constant over short periods, as the frames aligning to each state are modeled with static density

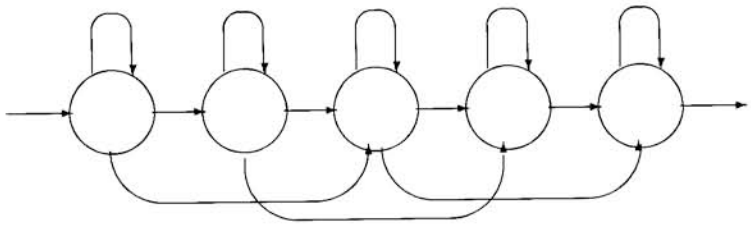

Figure 1: HMM topology for a single phoneme. Words are modeled by concatenating the models for the constituent phonemes.

functions. Going to an intermediate, linguistic feature representation, on the other hand, allows for a relaxation of this constraint. In particular, the global processing stage of our algorithm described in section ??, looks only at transitional regions in assessing the linguistic feature content of the signal. Thus it assumes that the abstract features, but not necessarily the physical ones, are piecewise constant.

In contrast to the physical representation described in this section, we develop in section ?? an intermediate interface between the physical and lexical representations of speech.

2.2 An Intermediate, Linguistic Feature Representation

The set of features which distinguish English phonemes is not unique; several sets have been introduced in the literature. The set which we shall adopt is, for the most part, that of Chomsky and Halle [?]. Specifically, we consider the following linguistic features, defined in terms of the required actions of a speaker in producing that sound and accompanied by specific spectral characteristics:

VOCALIC Sounds produced with an unconstricted oral cavity and with vocal cords which are positioned so as to allow spontaneous voicing. Vocalic sounds are typically loud in relation to non-vocalic sounds and exhibit visible formants (vocal-tract resonant frequencies) in the spectrum.

CONSONANTAL Includes sounds produced by forming an obstruction in the midsagittal region of the vocal tract, resulting in a lower total energy and lower first formant than non-consonantal sounds.

HIGH Sounds produced with the tongue body near the palate, resulting in a lowered first formant.

LOW Sounds produced with the tongue and jaw lowered, resulting in a high first formant.

BACK Includes those sounds produce with the tongue body toward the back of the mouth, resulting in a lowered second formant.

ANTERIOR Sounds produced with a constriction of the vocal tract anterior to the alveolar ridge.

CORONAL Includes those sounds for which the tongue blade is raised.

ROUND Sounds produced with rounded lips, causing all formants to lower in frequency.

TENSE Sounds produced with a deliberate and accurate gesture. Tense sounds are typically longer in duration with more extreme formant positions than non-tense sounds.

VOICE Sounds produced with the vocal folds vibrating, causing spectral resonances to become visible.

CONTINUANT Includes sounds for which the primary constriction of the vocal tract is not so narrow as to block the air flow past it, resulting in a smooth transition between the spectra associated with its predecessor and the spectra representing a continuant sound.

NASAL Sounds produced with the velum open. For nasal consonants the second formant is low in intensity and formant bandwidths are wide.

STRIDENT The air stream is directed against an obstructing surface, resulting in a noisy spectrum with substantial high-frequency energy.

LABIAL The primary constriction is formed at the lips, leading to a lowered first and second formant.

Tables ?? through ?? depict the binary linguistic feature representation of each of the vowels and the consonants we model. We refer to phonemes by the typewritten symbols used for labeling the TIMIT database. The "+" and "-" entries in the tables indicate the state of the corresponding articulator in the production of the sound. For example, sounds which are formed by rounding the lips are "+round" while sounds which do not involve lip rounding are "-round." Note that diphthongs have been excluded from the set of phonemes, as they consist of a transition from one feature vector to another, and therefore are the concatenation of two phonemes in this representation. In addition, neutral vowels have been omitted as the feature configurations for these sounds are volatile and "H" has been excluded because of its difficulty in fitting into the linguistic feature framework [?]. We include a representation of quiet in order to represent full closures in the same manner as the phonemes. We follow the

TIMIT notation of treating stop gaps as separate entities from the release even though linguistically these two units together comprise a single phoneme.

	VOWELS										
	IY	UW	EY	OW	AA	IH	UH	EH	AH	AO	AE
VOCALIC	+	+	+	+	+	+	+	+	+	+	+
CONSONANTAL	-	-	-	-	-	-	-	-	-	-	-
HIGH	+	+	-	-	-	+	+	-	-	-	-
BACK	-	+	-	+	+	-	+	-	+	+	-
LOW	-	-	-	+	+	-	-	-	-	-	+
ANTERIOR	-	-	-	-	-	-	-	-	-	-	-
CORONAL	-	-	-	-	-	-	-	-	-	-	-
ROUND	-	+	-	+	-	-	+	-	-	+	-
TENSE	+	+	+	+	+	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	+	+	+	+
CONTINUANT	+	+	+	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	-	-	-	-	-
LABIAL	-	-	-	-	-	-	-	-	-	-	-

Table 2: Linguistic features for each of the vowel sounds considered.

	GLIDES		LIQUIDS		NASALS			AFFRICATES		QUIET
	Y	W	L	R	M	N	NG	CH	JH	H#
VOCALIC	-	-	+	+	-	-	-	-	-	-
CONSONANTAL	-	-	+	+	+	+	+	+	+	+
HIGH	+	+	-	-	-	-	+	+	+	-
BACK	-	+	-	-	-	-	+	-	-	-
LOW	-	-	-	-	-	-	-	-	-	-
ANTERIOR	-	-	+	-	+	+	-	-	-	-
CORONAL	-	-	+	+	-	+	-	+	+	-
ROUND	-	+	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	-	+	-
CONTINUANT	+	+	+	+	-	-	-	-	-	+
NASAL	-	-	-	-	+	+	+	-	-	-
STRIDENT	-	-	-	-	-	-	-	+	+	-
LABIAL	-	-	-	-	+	-	-	-	-	-

Table 3: Linguistic features for each of the glide, liquid, nasal, and affricate phonemes considered, as well as the linguistic feature description of quiet.

	PLOSIVES						FRICATIVES							
	P	B	G	T	D	K	F	V	TH	DH	S	Z	SH	ZH
VOCALIC	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CONSONANTAL	+	+	+	+	+	+	+	+	+	+	+	+	+	+
HIGH	-	-	+	-	-	+	-	-	-	-	-	-	+	+
BACK	-	-	+	-	-	+	-	-	-	-	-	-	-	-
LOW	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ANTERIOR	+	+	-	+	+	-	+	+	+	+	+	+	-	-
CORONAL	-	-	-	+	+	-	-	-	+	+	+	+	+	+
ROUND	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VOICE	-	+	+	-	+	-	-	+	-	+	-	+	-	+
CONTINUANT	-	-	-	-	-	-	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	+	+	-	-	+	+	+	+
LABIAL	+	+	-	-	-	-	+	+	-	-	-	-	-	-

Table 4: Linguistic features for each of the plosive and fricative sounds considered.

Modern linguistic theory has departed from the notion of each phoneme being represented by the entire set of features. For example, since the production of vowels does not involve blocking the air flow through the vocal tract, the use of the feature continuant to describe vowels is unnecessary. The reduction of the representation to the non-redundant features describing each phoneme is efficient for the purposes of coding. However, from the viewpoint of recognition, the redundancies are desirable for recovery from errors as well as algorithm simplicity. We include the full set of feature descriptors for each phoneme as a sort of place keeper which will allow mathematical manipulation of our results, in much the same way that vectors lying in the x-y plane are specified as $[x, y, 0]$ in three dimensions.

Alternatives to the notion of "feature bundles," which connotes a lack of structure of the features, have been explored [?], [?]. Studies in feature geometries attempt to define hierarchical structures whose terminal nodes are the distinctive features.

In that spirit, we classify features as being primary or secondary:

- *Primary features:* Sonorant, Vocalic, Consonantal, Instant Release, Continuant
- *Secondary features:* Strident, Tense, High, Back, Low, Anterior, Coronal, Labial, Voice, Round, Nasal

where we have included in the list some commonly-discussed features which are not included in the inventory of Chomsky and Halle. Our algorithm embodies

a two-stage hierarchical graph, with secondary features conditional upon primary feature configurations. This is similar conceptually to imposing a Markov random field structure on the features in which a neighborhood is defined as a primary feature configuration.

We contend that the primary features determine the gross spectral characteristics of the resulting speech waveform; the other features modulate or make fine structural changes to the basic pattern defined by the primary ones. Therefore sounds which are characterized by the same primary features have similar spectra qualitatively, while sounds which have different primary features are fundamentally different. This fact implies that the features are encoded in the waveform hierarchically, with the manifestation of secondary features dependent on the configuration of the primary ones. For example, because G and IY have different configurations of primary features, the feature +high will be encoded differently in the waveform for the two phonemes.

The two-stage hierarchical search for features which is described in section ?? is essentially a search first for the manifestations of encoding a set of primary features in the neighborhood of each frame and then, given those features, a search for the secondary features, as well as a verification of the primary ones. The estimation of the broad class as a whole, as will be described, is meant to capture dependencies among the primary features.

3 Algorithm for Waveform Representation

In this section we describe the algorithm which results in a linguistic feature representation of speech waveforms. Initially we represent the waveform in terms of cepstra and their derivatives. The final representation is a probability vector at each frame; each component of the vector denotes the probability of a particular linguistic feature being encoded in the neighborhood of that frame.

An overview of the procedure is shown in figure ??. The initial stage of the hierarchical processing estimates the broad class of speech sounds represented. Based upon this estimate, we make both local and global inquiries as to the nature of the feature composition in the neighborhood of each frame. The terms local and global are chosen to emphasize that probabilities of features for a given frame are derived from narrow as well as wider windows in time around that frame. The outputs of the two levels of processing are averaged in order to arrive at the final estimate of the probability of each feature being encoded in the neighborhood of each frame.

Section ?? describes the broad class estimation stage of processing. Section ?? describes the temporally-local processing scheme by which we assign probabilities of features being encoded in the waveform in the neighborhood of each frame, while section ?? describes the temporally-global stage of processing.

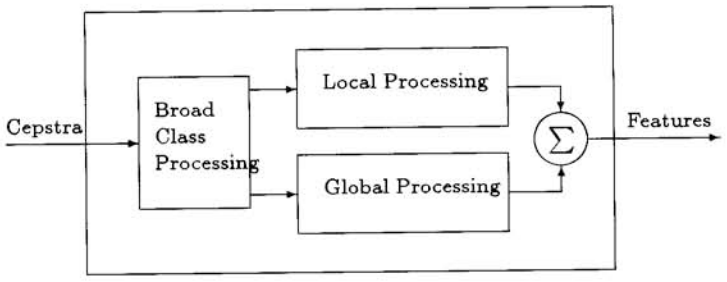

Figure 2: Overview of the system for assessing the linguistic feature content of a speech waveform.

3.1 Broad Class Modeling

The acoustic manifestation of a linguistic feature is dependent on the broad class of speech sounds to which the principally-represented phoneme belongs. For example, as mentioned in section ??, the feature +high will be encoded in the speech waveform differently in the cases of the vowel IY and the plosive G. Furthermore, all frames of a phone which corresponds to the presence or absence of a feature need not be spectrally similar. To capture time variations of the acoustic correlates of the features, we model separately the beginning, middle, and ending portions of the phones representing each broad class.

Each time frame t in the training set is assigned a truth label, $\tau(t) \in \{1, \ldots, 8\}$, reflecting the broad class of speech sounds which that frame represents. Our experiments are done on the TIMIT database which provides phonetic time markings. The broad-class label is the result of a many-to-few mapping of the TIMIT labels, as indicated in table ??.

TYPE OF SOUND	CONSTITUENT TIMIT LABELS
VOWEL	IY, UW, EY, OW, AA, IH, UH, EH, AH, AO, AE
GLIDE	Y, W
LIQUID	L, EL, R, ER, AXR
NASAL	M, EM, N, EN, NG
PLOSIVE	P, B, G, T, D, K
AFFRICATE	CH, JH
FRICATIVE	F, V, TH, DH, S, Z, SH, ZH
QUIET/VOICE BAR	H#, PCL, TCL, KCL, BCL, GCL, DCL, EPI, PAU

Table 5: Mapping from TIMIT label to broad class label.

Furthermore, each frame t of the training set is assigned a section label,

$\sigma(t) \in \{1, 2, 3\}$, indicating whether the frame represents the initial, middle, or final third of a phone. Each phone is divided into three pieces of equal duration in order to enable modeling of the time variation of the manifestation of a feature within each broad class.

The broad class and section labels are used to segregate the frames in the training set for the purposes of parameterizing statistical models. We have chosen unimodal Gaussians to model each broad class portion. However, more sophisticated models such as Gaussian mixtures or neural networks may be easily substituted. For section $i \in \{1, 2, 3\}$ of phones representing broad class $k \in \{1, \ldots, 8\}$ we estimate the model parameters as follows:

$$
\begin{aligned}
N_{k_i} &= \sum_{\{t|\sigma(t)=i,\tau(t)=k\}} 1 \\
\hat{\mu}_{k_i} &= \frac{1}{N_{k_i}} \sum_{\{t|\sigma(t)=i,\tau(t)=k\}} x(t) \\
\hat{\Sigma}_{k_i} &= \frac{1}{N_{k_i}} \sum_{\{t|\sigma(t)=i,\tau(t)=k\}} x(t)\, x^T(t) \; - \; \hat{\mu}_{k_i} \hat{\mu}_{k_i}^T
\end{aligned}
$$

That is, N_{k_i} is the total number of frames in the training set estimated as being drawn from the i-th third of a phone representing a phoneme in broad class k, $\hat{\mu}_{k_i}$ is the sample average (vector) over those same frames, and $\hat{\Sigma}_{k_i}$ is the sample covariance of that set of frames. We have that $N_{k_1} \approx N_{k_2} \approx N_{k_3}$ with differences arising due only to roundoff errors in dividing phones into thirds.

Given that section i of a phone representing broad class k is being produced, the $N-$dimensional probability density function for observations is taken as:

$$
p(x|\tau = k, \sigma = i) = \frac{1}{(2\pi)^{\frac{N}{2}} |\hat{\Sigma}_{k_i}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\hat{\mu}_{k_i})^T \hat{\Sigma}_{k_i}^{-1} (x-\hat{\mu}_{k_i})} \tag{1}
$$

The probability of each broad class being represented by each frame is calculated according to equation ??. These values will be used in the local processing scheme described in section ??. For the global processing scheme of section ?? we need to go a step further and estimate the sequence of broad class portions which each frame represents in a particular sentence. We assume a 3-state left-to-right Markov model of broad classes to find the most likely broad class sequence through an utterance. To estimate the transition probabilities among states, the number of transitions on a frame-by-frame basis among the broad class and section labels from the training data are counted. If we define the number of transitions from state s_i to state t_j in the training set as $T_{s_i t_j}$, then the transition probability between these two states is estimated as the number of transitions from state s_i to state t_j divided by the total number of transitions from state s_i:

$$
\hat{\alpha}_{s_i t_j} = \frac{T_{s_i t_j}}{\sum_{t=1}^{8} \sum_{j=1}^{3} T_{s_i t_j}}
$$

We assign initial state probabilities as:

$$\pi_{s_i} = \begin{cases} 1 & s = 8 \text{ and } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Dynamic programming is used to find the most likely sequence of broad classes arising in each sentence:

$$S_0{}^*, \ldots, S_M{}^* = \arg \max_{S_0, \ldots, S_M} \pi_{S_0} \prod_{m=1}^{M} p(x(t) \mid S_m)\, \hat{\alpha}_{S_{m-1}\, S_m}$$

where $S_m \in \{s_i : s \in \{1, \ldots, 8\}, i \in \{1, 2, 3\}\} \ \forall\ m$.

3.2 Local Processing

The clustering provided by the broad class labels is used in building models for frames which represent a given linguistic feature being encoded in the waveform as well as for frames which represent the absence of that feature. Again we have chosen unimodal Gaussian models, but more sophisticated statistical techniques may be incorporated in the general framework. Each frame $x(t)$ of the training set is assigned a label $\tau^f(t) \in \{0, 1\}$ indicating whether that frame corresponds to a "-" value ($\tau^f = 0$) or a "+" value ($\tau^f = 1$) of each linguistic feature $f \in \{1, \ldots, 14\}$. All frames in the TIMIT training set in a given portion of a broad class are divided into "feature +" and "feature -" subgroups for each linguistic feature to be modeled.

Gaussian models of the waveform attribute vectors are parameterized for each subgroup. We estimate the following model parameters:

$$\begin{aligned}
N_{k_i}^{f^+} &= \sum_{\{t \mid \sigma(t)=i, \tau(t)=k, \tau^f(t)=1\}} 1 \\
N_{k_i}^{f^-} &= \sum_{\{t \mid \sigma(t)=i, \tau(t)=k, \tau^f(t)=0\}} 1 \\
\hat{\mu}_{k_i}^{f^+} &= \frac{1}{N_{k_i}^{f^+}} \sum_{\{t \mid \sigma(t)=i, \tau(t)=k, \tau^f(t)=1\}} x(t) \\
\hat{\mu}_{k_i}^{f^-} &= \frac{1}{N_{k_i}^{f^-}} \sum_{\{t \mid \sigma(t)=i, \tau(t)=k, \tau^f(t)=0\}} x(t) \\
\hat{\Sigma}_{k_i}^{f^+} &= \frac{1}{N_{k_i}^{f^+}} \sum_{\{t \mid \sigma(t)=i, \tau(t)=k, \tau^f(t)=1\}} x(t)\ x^T(t) - \hat{\mu}_{k_i}^{f^+}\, \hat{\mu}_{k_i}^{f^+\ T} \\
\hat{\Sigma}_{k_i}^{f^-} &= \frac{1}{N_{k_i}^{f^-}} \sum_{\{t \mid \sigma(t)=i, \tau(t)=k, \tau^f(t)=0\}} x(t)\ x^T(t) - \hat{\mu}_{k_i}^{f^-}\, \hat{\mu}_{k_i}^{f^-\ T}
\end{aligned}$$

where $N_{k_i}^{f^+}$ indicates the number of frames which belong to portion i of a phone representing broad class k and which correspond to a "+" value of feature f. $\hat{\mu}_{k_i}^{f^+}$ and $\hat{\Sigma}_{k_i}^{f^+}$ represent the sample mean and sample covariance, respectively, of this set of frames. Similarly, $N_{k_i}^{f^-}$ indicates the number of frames which belong to portion i of a phone representing broad class k and which correspond to a "-" value of feature f, with $\hat{\mu}_{k_i}^{f^-}$ and $\hat{\Sigma}_{k_i}^{f^-}$ representing the sample mean and sample covariance of these frames.

The $N-$dimensional probability density function of frames which represent the presence of a given feature for portion i of a phone representing broad class k is modeled as:

$$p(x|k_i, \tau^f = 1) = \frac{1}{(2\pi)^{\frac{N}{2}} |\hat{\Sigma}_{k_i}^{f^+}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\hat{\mu}_{k_i}^{f^+})^T \hat{\Sigma}_{k_i}^{f^+ \ -1}(x-\hat{\mu}_{k_i}^{f^+})}$$

Similarly, the density function of frames which represent the absence of a given feature within the broad class portion is modeled as:

$$p(x|k_i, \tau^f = 0) = \frac{1}{(2\pi)^{\frac{N}{2}} |\hat{\Sigma}_{k_i}^{f^-}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\hat{\mu}_{k_i}^{f^-})^T \hat{\Sigma}_{k_i}^{f^- \ -1}(x-\hat{\mu}_{k_i}^{f^-})}$$

We use Bayes' Rule to give the probability of a given feature being encoded in the neighborhood of a given frame:

$$\begin{aligned} p(\tau^f = 1|x)p(x) &= \sum_{i,k} p(x|k_i, \tau^f = 1)p(\tau^f = 1|k_i)p(k_i|x) \\ p(\tau^f = 0|x)p(x) &= \sum_{i,k} p(x|k_i, \tau^f = 0)p(\tau^f = 0|k_i)p(k_i|x) \\ Pr(\tau^f = 1|x) &= \frac{p(\tau^f = 1|x)}{p(\tau^f = 1|x) + p(\tau^f = 0|x)} \end{aligned}$$

The local processing algorithm has some shortcomings which are addressed through the complementary global processing of the next section. The primary weakness of the local processing is that decisions about the feature composition of each frame are made based upon a small neighborhood around that frame. Information about spectra elsewhere in the waveform is incorporated into the attribute vector describing each frame only through the use of cepstral derivatives. A second weakness of the local processing is its failure to fully model the interaction among features within each phoneme. Features are modeled within estimated broad class regions, thereby modeling the dependence of the secondary features on the configuration of primary features. However, correlations among secondary features are lost.

3.3 Global Processing

The weaknesses of the the local processing discussed in section ?? are addressed through the complementary global processing described in this section. The global processing takes into account the information present in the spectra associated with one sound in describing the feature composition of a neighboring sound by explicitly modeling transitional regions, resulting in a more global process than the procedure described in section ??. In contrast to the bottom-up procedure of the local processing, in which frames representing the "+" and "-" values of each feature were modeled directly within each broad class portion, we use a top-down procedure in this section. Models here are at the diphone level, and a mapping is performed to transform probabilities of phones into probabilities of features. This method captures the potential interdependence of features by modeling entire feature sets, or phonemes.

The coarticulatory effects introduced in continuous speech warrant processing at a global level in order to incorporate information about a sound which appears in the spectra outside of the corresponding phone as well as the deformation of that phone due to feature spread across transitional regions.

The global algorithm relies on models of spectra in the neighborhood of transitions. We shun the notion of explicitly segmenting the waveform or searching for "landmarks" as a preprocessing technique. Rather, we have viewed segmentation as a byproduct of recognition on a course scale; a change in a broad class estimate implies that a transition from one sound to another has occurred. This approach is motivated by a view of segmentation as a phenomenon on an abstract level rather than a physical one. The method for inferring a segmentation is able to incorporate durational cues directly through the transition probabilities in the Markov model.

We construct an attribute vector $\tilde{x}$ for a transition occurring at time t, where

$$\tilde{x}(t) = \begin{bmatrix} \sum_{k=0}^{2} x(t-k) \\ \\ \sum_{k=1}^{3} x(t+k) \end{bmatrix}$$

and x is a 42-dimensional vector consisting of 14 normalized cepstra and their first and second time derivatives.

$\mathcal{T}$ is defined to be the set of transitions in the data set; for training we take $\mathcal{T} = \{t : \phi(t) \neq \phi(t+1)\}$, where $\phi(t)$ is the TIMIT label occurring at frame t. For testing, we take $\mathcal{T} = \{t : \hat{\sigma}(t) = 3, \hat{\sigma}(t+1) = 1\}$, where $\hat{\sigma}(t) \in \{1, 2, 3\}$ is the portion of a broad class at time t.

The limited number of training tokens of each transition mandates the use of reduced-order models. We have used linear-discriminant analysis to reduce the original 84 dimensions to 25. This technique consists of defining a set of classes and computing the "within-class" and "between-class" covariance matrices of the training samples. The within-class matrix W is defined as

$E[(\tilde{x} - \mu_{c_{\tilde{x}}})(\tilde{x} - \mu_{c_{\tilde{x}}})^T]$ where $c_{\tilde{x}}$ is the class of observation $\tilde{x}$ and $\mu_{c_{\tilde{x}}}$ is the mean of that class; classes for the global processing LDA are phoneme transition pairs. The between-class covariance matrix B is defined as $E[(y - \mu)(y - \mu)^T]$ where μ is the average across all the training data. The eigenvectors associated with the N largest eigenvalues of $W^{-1}B$ are taken as the N rows of the dimension-reduction matrix A.

For each transition in testing, we perform the transformation $x' = A\tilde{x}$ and evaluate the Gaussian models for each possible phone transition. The following quantities are calculated for transitions from phoneme α to phoneme β for each $\alpha \neq \beta$:

$$\begin{aligned}
\mu'_{\alpha\beta} &= \frac{1}{N_{\alpha\beta}} \sum_{\{t|\phi(t)=\alpha,\phi(t+1)=\beta\}} x'(t) = A\mu_{\alpha\beta} \\
\Sigma'_{\alpha\beta} &= \frac{1}{N_{\alpha\beta}} \sum_{\{t|\phi(t)=\alpha,\phi(t+1)=\beta\}} x'(t)x'(t)^T - \mu'_{\alpha\beta}{\mu'_{\alpha\beta}}^T = A\Sigma_{\alpha\beta}A^T
\end{aligned}$$

We shall say that $\alpha \ni f$ if the feature f is specified as "+" in the vector of features representing phoneme α. We have that the probability of feature f being encoded in the waveform in the left neighborhood of a transition at time t is:

$$\begin{aligned}
p_t(f|x') &= \sum_{\alpha \ni f} p(\phi(t) = \alpha|x') \\
&= \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(\phi(t) = \alpha, \phi(t+1) = \beta|x') \\
&= \frac{1}{p(x')} \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(x'|\phi(t) = \alpha, \phi(t+1) = \beta)p(\phi(t) = \alpha, \phi(t+1) = \beta)
\end{aligned}$$

This procedure sums over all possible right contexts for a given phone, and then sums over all phones in which a given feature has the value "+" in order to arrive at the probability that a feature is encoded in the neighborhood to the left of the transition. We can also assess this probability for the frames to the right of a transition by summing over left contexts. We have that the probability of feature f being encoded in the waveform in the right neighborhood of a transition at time r is:

$$\begin{aligned}
p_r(f|x') &= \sum_{\alpha \ni f} p(\phi(r+1) = \alpha|x') \\
&= \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(\phi(r) = \beta, \phi(r+1) = \alpha|x') \\
&= \frac{1}{p(x')} \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(x'|\phi(r) = \beta, \phi(r+1) = \alpha)p(\phi(r) = \beta, \phi(r+1) = \alpha)
\end{aligned}$$

The average of the probabilities derived from the left and right transitions provides a globally-derived estimate of the probability of feature f being encoded between the two transitions. That is, $\forall s : r \leq s < t$, $p_s(f|x') = \frac{1}{2}(p_{r+}(f|x') + p_{t-}(f|x'))$.

In order to combine this global estimate of the probability of a given feature being encoded in the waveform in the neighborhood of each frame with the local estimate described in section ?? the average of the probabilities derived from these two processes is taken.

4 Lexical Access by Features

The benefit of the feature representation lies in its enabling of a linguistically-motivated method of lexical access. We have viewed the task of assessing confidence in the recognition of a particular word of interest as an opportunity to perform lexical access on the basis of linguistic features.

In our paradigm, a continuous speech recognition system finds the most likely word sequence to account for an utterance. In addition, a set of keywords is chosen. Each time one of the keywords is hypothesized, the event is labeled with a score which is the average likelihood of the frames along the recognition path of the HMM which pass through the keyword. This HMM score can be interpreted as confidence in the hypothesis. From a receiver point of view, hypothesizing a word of interest when that word was uttered corresponds to a detection while incorrectly hypothesizing the keyword corresponds to a false alarm. Thus, by ordering the HMM scores corresponding to the hypothesized occurrences of a keyword we can generate a receiver operating characteristic (ROC) to characterize our ability to efficiently detect that word.

In this section we consider the detection of the keyword "cuatro" in a Spanish recognition task. Each utterance in which that word was hypothesized was represented in terms of its linguistic feature content. We then used the state alignment produced by the HMM to compare the observed linguistic features to the configurations expected for the keyword for all frames aligning to states within the word of interest.

Each phoneme of the keyword represents a binary linguistic feature configuration, while each observation is represented by a real-valued feature vector. For each phoneme in the keyword we compute the average L_1 distance to its theoretical configuration of all of the observation frames which aligned to that phoneme. The individual phoneme scores are then averaged to generate an overall score for the hypothesized occurrence of the keyword.

We compare the ROC curve generated by ranking the HMM scores for each hypothesized occurrence of "cuatro" with that generated by ranking the linguistic feature scores for the same events. The dashed line in figure ??, corresponding to the HMM score, is little better than random; the ROC associated with the linguistic feature score, shown in the solid line, indicates a significant

Figure 3: ROC curves for the word "cuatro." Dashed line indicates HMM scoring, while solid line corresponds to linguistic feature scoring.

increase in detection performance.

5 Conclusions and Future Directions

In this paper we have described the general framework which we use to parameterize the speech waveform in terms of linguistic features. We emphasize that the specific models we chose could be increased in complexity as the amount of training data increases if speed or storage constraints allow. In particular, the use of Gaussian mixture models in the broad class and local processing stages would enable sharper modeling of the decision spaces. However, the fact that we achieve good results with simple unimodal Gaussian models indicates that phonemically relevant aspects of the signal are indeed being modeled.

While our algorithm is a feed-forward one, we envision closing the feedback loop. The estimated features can be compared with theoretical configurations. The trajectory of the deviations from the ideal configurations may provide some

information about the speaker. Compensation for the speaker-dependent aspects of the waveform by adjusting the input cepstra should reduce the variability in the input signal and make the modeling task easier, thereby increasing its accuracy.

The linguistic feature representation provides an intermediate abstraction appropriate for lexical access. We have described our approach to the task of keyword detection from recognition, in which we perform lexical access on the basis of features. We have shown the linguistic feature approach to provide better detection for a given false alarm rate than what is achieved by ordering the scores provided by the HMM.

References

[1] Chomsky, N. and M. Halle. *Sound Pattern of English.* New York: Harper & Row. 1968.

[2] Clements, G. "The Geometry of Phonological Features." in *Phonology Yearbook 2.* 1985. pp 225-252.

[3] Fant, G. *Speech Sounds and Features.* Cambridge, MA: MIT Press. 1973.

[4] McCarthy, J. "Feature Geometry and Dependency: A Review." in *Phonetica 43.* 1988; 45:84-108.

[5] Papoulis, A. *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill. 1984. Cliffs, NJ: Prentice-Hall. 1987.

[6] Sagey, E. *The Representation of Features and Relations in Non-linear Phonology.* PhD Dissertation. Massachusetts Institute of Technology. 1986.

[7] Stevens, K. "Evidence for the Role of Acoustic Boundaries in the Perception of Speech Sounds" in *Phonetic Linguistics. Essays in Honor of Peter Ladefoged.* V. Fromkin, ed. Orlando, FL: Academic Press, Inc. 1985.

[8] Stevens, K. "Phonetic Features and Lexical Access." Presented at em Symposium on Advanced Man-Machine Interface Through Spoken Language. November 19-22, 1988. Hawaii.

[9] Stevens, K. and M. Halle. "Remarks On Analysis By Synthesis and Distinctive Features." in *Models for the Perception of Speech and Visual Form.* Cambridge, MA: The MIT Press. 1964.

[10] J.R. Rohlicek, W. Russell, S. Roucos, and H. Gish, "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting," in *IEEE ICASSP* 1989, pp. 627-630.