A LINGUISTIC FEATURE REPRESENTATION OF THE SPEECH WAVEFORM

Ellen Eide[1,2] *J. Robin Rohlicek*[1] *Herbert Gish*[1] *Sanjoy Mitter*[2]

[1]BBN Systems and Technologies
70 Fawcett Street 15/1c
Cambridge, MA 02138 USA

[2]Massachusetts Institute of Technology
77 Massachusetts Avenue 35-313
Cambridge, MA 02139 USA

ABSTRACT

Linguistic theory views a phoneme as a shorthand notation for a bundle of binary features related to the operation of the speaker's articulators. In this paper, a representation of the speech waveform in terms of these underlying distinctive features is described. The estimation of the probability of each of fourteen linguistic features being encoded locally in the waveform is performed on a frame-by-frame basis. In going from the abstract to the physical level, we recognize that the features are encoded in the waveform hierarchically and that time-varying manifestations of a feature within a phonemic segment are possible. These issues are addressed simultaneously through a two-stage procedure. In the first pass, the time portion and broad class of sound being represented by each frame is estimated. On the second pass, for each distinctive linguistic feature, models built explicitly for the estimated broad class portion are evaluated to arrive at the probability that each frame is part of a realization of a phoneme in which the feature is present. The distinctive feature representation is applied to the tasks of phoneme recognition and secondary classification in keyword spotting. The wordspotting algorithm compares estimated linguistic feature vectors to idealized target configurations.

1. LINGUISTIC FEATURE REPRESENTATION

The goal of this work is to adapt an abstract, linguistic feature representation of speech to a representation at the waveform level. Tables 1 through 4 depict the binary linguistic feature representation of each of the vowels and the consonants distinguished in this study.

	VOWELS										
	IY	UW	EY	OW	AA	IH	UH	EH	AH	AO	AE
VOCAL.	+	+	+	+	+	+	+	+	+	+	+
CONS.	-	-	-	-	-	-	-	-	-	-	-
HIGH	+	+	-	-	-	+	+	-	-	-	-
BACK	-	+	-	+	+	-	+	-	+	+	-
LOW	-	-	-	+	+	-	-	-	-	-	+
ANTER.	-	-	-	-	-	-	-	-	-	-	-
CORON.	-	-	-	-	-	-	-	-	-	-	-
ROUND	-	+	-	+	-	-	+	-	-	+	-
TENSE	+	+	+	+	+	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	+	+	+	+
CONT.	+	+	+	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-
STRID.	-	-	-	-	-	-	-	-	-	-	-
LABIAL	-	-	-	-	-	-	-	-	-	-	-

Table 1. Feature representation of each of the vowels considered.

Acoustic manifestations of a feature for a given phoneme are dependent on the broad class of speech sounds to which that phoneme belongs. This is equivalent to stating that features are encoded hierarchically, as the broad class may be determined from a subset of the features. Because of this hierarchical structure, we perform a two-stage analysis, wherein the goal of the first stage is to provide an estimate of the broad class of sounds represented by each frame. Processing in the second stage relies upon this estimate. In addition, because the fact that each phoneme is characterized by a single vector of features does not imply that all frames of a given phone share roughly the same spectra, we model separately the onset, middle, and end of each broad class.

Feature analysis of a waveform results in the parameterization of each frame by the probabilities that each of the distinctive features is encoded locally in the waveform.

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

Table 2. Feature representation of each of the glides, liquids, nasals, and affricates considered.

The estimation of the broad class portion represented by each frame relies upon a set of Gaussian models whose parameters are derived from a TIMIT training set. We have used the data from all male speakers in the northern and north midland dialects, as well as speakers ADD through PAB of the western dialect for training. Western male speakers PAR through WRP form our TIMIT testing set. Waveforms have a bandwidth of 0-8 kHz, a frame rate of 5ms, and an analysis window of 15ms. Each frame in the training set is assigned a truth label indicating that it represents either the beginning, middle or final third of a fricative, nasal, vowel, liquid, glide, quiet/closure, or instant release segment. The mean and covariance of all attribute vectors in the training set representing a given broad class portion are computed, where the attribute vectors consist of normalized cepstra $NC_0 - NC_{13}$ and derivative cepstra $DC_0 - DC_{12}$.

In order to find the most-likely sequence of broad classes in an utterance, the probabilities of each broad class occurring at each time frame are treated as observations from a Markov source,

0-7803-0946-4/93 $3.00

	PLOSIVES						SILENCE
	P	B	G	T	D	K	H#
VOCALIC	-	-	-	-	-	-	-
CONSONANTAL	+	+	+	+	+	+	+
HIGH	-	-	+	-	-	+	-
BACK	-	-	+	-	-	+	-
LOW	-	-	-	-	-	-	-
ANTERIOR	+	+	-	+	+	-	-
CORONAL	-	-	-	+	+	-	-
ROUND	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-
VOICE	-	+	+	-	+	-	-
CONTINUANT	-	-	-	-	-	-	+
NASAL	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	-
LABIAL	+	+	-	-	-	-	-

Table 3. Feature representation of each of the plosives considered, and of silence.

with each state in the Markov model corresponding to a portion of a broad speech class. The number of transitions on a frame-by-frame basis between the broad class truth labels in the training data are counted to estimate state transition probabilities. Finally, dynamic programming is used to find the most-likely broad class sequence in each sentence.

	FRICATIVES							
	F	V	TH	DH	S	Z	SH	ZH
VOCALIC	-	-	-	-	-	-	-	-
CONSONANTAL	+	+	+	+	+	+	+	+
HIGH	-	-	-	-	-	-	+	+
BACK	-	-	-	-	-	-	-	-
LOW	-	-	-	-	-	-	-	-
ANTERIOR	+	+	+	+	+	+	-	-
CORONAL	-	-	+	+	+	+	+	+
ROUND	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-
VOICE	-	+	-	+	-	+	-	+
CONTINUANT	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-
STRIDENT	+	+	-	-	+	-	-	-
LABIAL	+	+	-	-	-	-	-	-

Table 4. Feature representation of each of the fricatives considered.

After the broad class portion represented by each frame is estimated, the probabilities of each of the linguistic features being represented are evaluated from Gaussian models built explicitly from training data *estimated* to have been representatives of that broad class portion.

In order to evaluate the parameters of the Gaussians, for each time frame in the training set, the TIMIT phoneme label is mapped to a set of fourteen binary truth values representing the linguistic feature configuration of the underlying phoneme. The clustering of the training data provided by the broad class estimation stage is used to build twenty-four sets of linguistic feature models. All frames in the TIMIT training set estimated to belong to a given broad class portion are divided into feature-present and feature-absent sets for each of the features and the mean and covariance of the waveform attribute vectors for each of these sets are calculated. In testing, the broad class portion estimate of each frame keys the choice of Gaussian models from which to evaluate the probabilities of each of the features being present.

In order to evaluate performance of the algorithm, a dynamic programming stage to decide the presence or absence of each feature is included. A measure of the algorithm's performance on an individual phoneme basis is given in figure 1. For each of the phonemes listed in the left-hand column, the relative frequency of the frames corresponding to that phoneme which were estimated to represent the *presence* of the features listed horizontally is given. The "+" or "-" following each entry indicates the theoretical presence or absence of the feature.

2. PHONEME IDENTIFICATION FROM FEATURE PROBABILITIES

Given a set of phonemic boundaries, the task at hand is to identify the phone occurring between the boundary points. The training and testing waveforms for this experiment are those described above. In deriving the broad class estimates, however, we make use of the known boundaries by forcing the estimated state sequence to change from the end of one broad class to the beginning of another only at the boundaries. Broad class portion estimates are allowed to change anywhere between boundaries.

Assuming a segment has duration T, we choose the estimate to be

$$\hat{\phi} = \text{argmax}_{\phi} p(\phi | f_1, \ldots, f_T, T)$$

where f_t is the vector of linguistic feature probabilities at time t.

To begin, we define $q_t : \Re^{14} \rightarrow Z^+ \cup \{0\}$ as:

$$q_t = \sum_{m \in \{1,\ldots,14 | f_m(t) > 0.5\}} 2^{m-1}$$

Thus, q_t serves as a quantizer for the real-valued feature probability vector f_t, so that $p(\phi|f) \approx p(\phi|q)$ where the latter term may be estimated as the number of occurrences of phoneme ϕ and index q relative to the total number of occurrences of index q. Assuming that the duration and observed features are independent,

$$\begin{aligned} \hat{\phi} &= \text{argmax}_{\phi} p(\phi | f_1, \ldots, f_T, T) \\ &= \text{argmax}_{\phi} p(f_1, \ldots, f_T, T | \phi) p(\phi) \\ &= \text{argmax}_{\phi} p(f_1, \ldots, f_T | \phi) p(\phi) p(T | \phi) \\ &= \text{argmax}_{\phi} p(\phi | f_1, \ldots, f_T) p(T | \phi) \\ &= \text{argmax}_{\phi} p(\phi | q_1, \ldots, q_T) p(T | \phi) \end{aligned}$$

The first term on the right-hand side of the above is approximated by:

$$P(\phi | q_1, \ldots, q_T) \approx \left(\prod_{t=1}^{T} p(\phi | q_t) \right)^{\frac{1}{T}}$$

Also, we model duration as a third order Erlang, so that:

$$p(T|\phi) = \alpha_{\phi}{}^{T-3} (1 - \alpha_{\phi})^3$$

where

$$\alpha_{\phi} = \frac{N_{\phi}}{3 + N_{\phi}}$$

and N_{ϕ} is the number of times a frame representing phoneme ϕ follows a frame representing the same phoneme in the training set.

A recognition rate of 70% has been achieved using the linguistic feature representation to estimate the phonemes in hand-marked segments of the TIMIT testing waveforms for the phonemes listed in tables 1 through 4. A rate of 63.5% was achieved within the submatrix of vowels listed in table 1.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTER.	CORON.	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.828+	0.108-	0.926+	0.018-	0.009-	0.046-	0.077-	0.004-	0.868+	0.981+	0.938+	0.066-	0.020-	0.019-
UW	0.812+	0.433-	0.376+	0.698+	0.313-	0.368-	0.393-	0.601+	0.544+	0.986+	0.829+	0.293-	0.105-	0.217-
EY	0.972+	0.035-	0.943-	0.062-	0.050-	0.025-	0.028-	0.014-	0.833+	0.985+	0.979+	0.016-	0.009-	0.010-
OW	0.971+	0.247-	0.093-	0.902+	0.770+	0.472-	0.319-	0.865+	0.453+	0.982+	0.980+	0.116-	0.005-	0.006-
AA	0.968+	0.074-	0.011-	0.905+	0.880+	0.079-	0.094-	0.565-	0.372+	0.982+	0.974+	0.008-	0.005-	0.014-
IH	0.909+	0.100-	0.802+	0.163-	0.111-	0.078-	0.076-	0.156-	0.269-	0.971+	0.951+	0.012-	0.015-	0.014-
UH	0.960+	0.225-	0.391+	0.715+	0.411-	0.166-	0.205-	0.775+	0.166-	0.980+	0.974+	0.020-	0.007-	0.020-
EH	0.952+	0.182-	0.570-	0.260-	0.381-	0.048-	0.153-	0.139-	0.095-	0.973+	0.959+	0.018-	0.011-	0.015-
AH	0.958+	0.091-	0.122-	0.847+	0.690-	0.135-	0.109-	0.449-	0.131-	0.984+	0.968+	0.049-	0.005-	0.019-
AO	0.973+	0.143-	0.011-	0.933+	0.727-	0.122-	0.132-	0.827+	0.384-	0.982+	0.986+	0.035-	0.009-	0.014-
AE	0.963+	0.034-	0.198-	0.165-	0.709+	0.018-	0.022-	0.008-	0.013-	0.987+	0.986+	0.003-	0.004-	0.007-
Y	0.488-	0.242-	0.848+	0.035-	0.344-	0.141-	0.141-	0.016-	0.453-	0.949+	0.840+	0.266-	0.008-	0.027-
W	0.847-	0.223-	0.419+	0.783+	0.683-	0.180-	0.149-	0.758+	0.046-	0.952+	0.955+	0.458-	0.008-	0.020-
L	0.882+	0.489+	0.109-	0.567-	0.392-	0.760+	0.567+	0.485-	0.299-	0.973+	0.933+	0.134-	0.014-	0.056-
R	0.927+	0.855+	0.087-	0.163-	0.130-	0.040-	0.820+	0.092-	0.068-	0.960+	0.937+	0.029-	0.055-	0.084-
N	0.165-	0.890+	0.182-	0.210-	0.048-	0.844+	0.847+	0.022-	0.024-	0.984+	0.271-	0.857+	0.032-	0.137-
M	0.103-	0.927+	0.047-	0.111-	0.035-	0.879+	0.611-	0.024-	0.017-	0.967+	0.220-	0.857+	0.093-	0.665+
NG	0.244-	0.781+	0.541+	0.571+	0.080-	0.691-	0.716-	0.038-	0.175-	0.992+	0.248-	0.811+	0.008-	0.114-
K	0.000-	1.000+	0.688+	0.722+	0.000-	0.090-	0.078-	0.000-	0.000-	0.016-	0.153-	0.001-	0.022-	0.024-
T	0.000-	0.999+	0.162-	0.026-	0.000-	0.717+	0.722+	0.000-	0.000-	0.042-	0.132-	0.002-	0.194-	0.044-
P	0.000-	1.000+	0.031-	0.080-	0.000-	0.758+	0.098-	0.000-	0.000-	0.065-	0.174-	0.000-	0.037-	0.714+
D	0.026-	0.972+	0.110-	0.017-	0.013-	0.656+	0.530+	0.000-	0.015-	0.489+	0.264-	0.052-	0.095-	0.102-
B	0.025-	0.986+	0.018-	0.018-	0.022-	0.652+	0.210-	0.011-	0.011-	0.877+	0.373-	0.069-	0.029-	0.732+
G	0.005-	1.000+	0.516+	0.579+	0.000-	0.240-	0.113-	0.000-	0.000-	0.543+	0.231-	0.032-	0.005-	0.045-
J	0.000-	1.000+	0.732+	0.003-	0.000-	0.268-	0.873+	0.000-	0.000-	0.386+	0.157-	0.009-	0.810+	0.000-
CH	0.000-	1.000+	0.879+	0.000-	0.000-	0.254-	0.938+	0.000-	0.000-	0.017-	0.092-	0.000-	0.902+	0.000-
F	0.030-	0.975+	0.016-	0.032-	0.011-	0.854+	0.151-	0.008-	0.011-	0.047-	0.829+	0.007-	0.799+	0.867+
S	0.028-	0.977+	0.027-	0.015-	0.009-	0.963+	0.971+	0.009-	0.010-	0.043-	0.951+	0.008-	0.952+	0.001-
SH	0.030-	0.974+	0.907+	0.003-	0.001-	0.127-	0.962+	0.003-	0.017-	0.037-	0.350+	0.001-	0.957+	0.000-
TH	0.030-	0.970+	0.045-	0.022-	0.013-	0.760+	0.515+	0.015-	0.011-	0.056-	0.794+	0.006-	0.569-	0.483-
DH	0.082-	0.950+	0.061-	0.065-	0.029-	0.659+	0.552+	0.021-	0.025-	0.517+	0.569+	0.142-	0.174-	0.169-
Z	0.049-	0.960+	0.066-	0.012-	0.007-	0.942+	0.958+	0.009-	0.016-	0.359+	0.938+	0.009-	0.929+	0.003-
ZH	0.028-	0.944+	0.611+	0.028-	0.028-	0.361-	0.944+	0.000-	0.028-	0.417+	0.056+	0.000-	0.944+	0.000-
V	0.242-	0.842+	0.087-	0.159-	0.110-	0.647+	0.328-	0.094-	0.048-	0.826+	0.799+	0.326-	0.558+	0.641+
H#	0.003-	0.999+	0.002-	0.001-	0.001-	0.007-	0.008-	0.000-	0.001-	0.006-	0.996+	0.002-	0.005-	0.001-

Figure 1. Performance of estimating features for individual phonemes when phonemic boundaries are unknown.

3. KEYWORD SPOTTING

In keyword spotting, no constraints on the speaker's vocabulary are imposed. The open set of allowable utterances poses a difficulty in modeling viable alternatives to the keywords. Representation of the speech waveform in terms of linguistic features, however, introduces a closed set of allowable feature configurations. The trajectory of the 14-dimensional feature vector traces a path through the feature space $[0,1]^{14}$, with a binary-valued target vector $\phi \in \Phi$ existing for each phoneme in the keyword, where Φ is the set of phonemes in the language. A keyword represents an ordered set of targets $\{\phi_j\}_{j=1}^{J} \in \mathcal{K} \subset \Phi$, or equivalently, an ordered set of corners in $[0,1]^{14}$.

The linguistic feature representation has been applied in the context of secondary classification. A list of putative occurrence locations and a score reflecting the likelihood that the keyword actually occurred is provided by BBN's hidden Markov model (HMM) [2]. The goal of the secondary processing is to improve the receiver operating characteristic (ROC) over that based upon the HMM score alone.

The set of putative occurrences of each keyword is used to extract from the conversation a small set of events to be scored. For each event, 14 normalized cepstral coefficients and 14 first derivatives are calculated at each time frame. TIMIT models as described in section 1, but trained on narrowband (300-3300 Hz) data, are used to provide an estimate of the probability of each linguistic feature being present at each time frame.

The normalized L_1 distance at frame t, $d(\phi, f_t)$, between a target, ϕ, and the vector of feature estimates, f_t, is computed for each frame in a putative occurrence of a given keyword:

$$d(\phi, f_t) = 1 - \frac{1}{14}\sum_{m=1}^{14} \phi_m f_{t_m} + (1-\phi_m)(1-f_{t_m})$$

KEYWORD	HMM FOM	HMM & Linguistics FOM
chester	66.8%	74.0%
conway	90.6	91.7
interstate	86.3	91.1
look	5.7	18.6
middleton	66.3	67.5
minus	67.9	77.0
mountain	60.1	71.4
road	46.9	47.4
thicket	72.6	88.6

Table 5. Figures of merit for each of the keywords subjected to secondary processing.

Each phoneme ϕ present in the phonetic spelling of a given keyword is represented by a collection $\mathcal{X}_\phi$ of three states in a left-to-right Markov chain. The state of the process at time t, x_t, is the state of the Markov chain with which observation t will be aligned. Self-transition probabilities α_{ii} are derived from TIMIT phoneme labels; interstate transition probabilities α_{ii+1} are taken as $1 - \alpha_{ii}$.

The control $u_t \in \{0,1\}$ keys the transition between adjacent states in the Markov chain; i.e. $x_{t+1} = x_t + u_t$. Distances $d(\phi, f_t)$, $\phi \in \mathcal{K}, t \in \{1,\ldots,T\}$, along with the transition probabilities α_{ij} define an incremental cost

$$C_{ij}(t) = -\beta\, d(\phi, f_t) + \log(\alpha_{ij}) + \log(\frac{\beta}{1-e^{-\beta}})$$

of setting $x_t = j \in \mathcal{X}_\phi$ given that $x_{t-1} = i$. β is a constant which controls the relative contributions of each of the terms to the cost function. Dynamic programming is used to find the controls $\{u_t\}$, or equivalently the state sequence P^*, which maximize the cost function, subject to the endpoint constraints $x_0 = 1$ and $x_T = 3J$.

Maximization of this cost function is equivalent to finding the

most-likely alignment of the data to the Markov chain when we model the L_1 distance of a phoneme to its ideal target as having an exponential distribution in $[0, 1]$ with decay rate β:

$$P(f_t|\phi) = \frac{\beta}{1 - e^{-\beta}} e^{-\beta d(f_t, \phi)}$$

The accumulated average cost of P^* through the target sequence is taken as the score σ_{LP} for the putative event:

$$\sigma_{LP} = \sum_{\phi_j \in \mathcal{K}} \frac{1}{N_{\phi_j}} \sum_{\{t | x_t \in \mathcal{X}_{\phi_j}\}} C_{x_t x_{t+1}}(t)$$

where N_{ϕ_j} is the total number of frames aligned with target ϕ_j.

We use the stonehenge database for training; the putative events from the HMM for speakers sm03c-sm16c provide a set of hits and false alarms from which we perform hand tuning of the parameter β for each keyword, as well as determine the phonetic spelling by which to represent the word.

The putative events from the HMM for the stonehenge speakers sm33c-sm43c form the set of testing data. We combine the linguistic feature score with the HMM score for each event linearly, using weights hand-tuned from the training set, to provide an overall score for each putative event. The set of overall scores, when ordered, provide a means of secondary classification. Shown in table 5 are the figures of merit (FOM), defined as the average probability of detection from 0 to 10 false alarms per keyword per hour, for each keyword submitted to secondary processing. To arrive at the score in the right hand column, a combination of the HMM score σ_{HMM} and the linguistic feature score of the form:

$$c_1\sigma_{LP} + c_2\sigma_{LP}\,\sigma_{HMM} + \sigma_{HMM}$$

was used, with the weights c_1 and c_2 determined from hand-tuning on the training set. As shown in the table, a significant increase in performance results from the secondary processing, indicating that the linguistic features extract information in the waveform differently from the HMM.

4. SUMMARY

We have devised a representation of the speech waveform in terms of a set of abstract linguistic features which pertain to the mode of operation of the articulators in producing a sound. The representation has been applied to the tasks of phoneme identification and secondary classification in wordspotting.

5. REFERENCES

1. Chomsky, N. and M. Halle. 1968. *Sound Pattern of English.*

2. 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.