KNOWLEDGE-BASED PARAMETERS FOR HMM SPEECH RECOGNITION

Nabil N. Bitar and Carol Y. Espy-Wilson

Electrical, Computer and Systems Engineering Department
Boston University, Boston, MA 02215

ABSTRACT

This paper presents acoustic parameters (AP's) that were motivated by phonetic feature theory and employed as a signal representation of speech in a Hidden Markov Model (HMM) recognition framework. Presently, the phonetic features considered are the manner features: sonorant, syllabic, nonsyllabic, noncontinuant and fricated. The objective of the parameters is to directly target the linguistic information in the signal and to reduce the speaker-dependent information that may yield large speech variability. To achieve these goals, the AP's were defined in a relational manner across time or frequency. For evaluation, broad-class recognition experiments were conducted comparing the AP's to cepstral-based parameters. The results of the experiments indicate that the AP's are able to capture the phonetically relevant information in the speech signal and that, in comparison to the cepstral-based parameters, they are more able to reduce the interspeaker variability.

1. INTRODUCTION

In our research, we seek acoustic parameters (AP's) that target the phonetically distinctive information in the speech signal and reduce speaker-dependent information (e.g. gender). The objective is to employ these acoustic parameters as a signal representation for speaker-independent speech recognition. Employing such parameters in the recognition process has two advantages: (1) it allows incorporation of speech knowledge (2) it serves as a tool for understanding acoustic-phonetics and contextual variability.

In order to achieve our goal, the AP's are (1) based on phonetic-features [1] and (2) defined in a relational manner. Phonetic features describe the manner and place of speech production. They are the minimal speech units needed to distinguish among the most similar sounds of language (e.g. the phonetic feature voiced distinguishes the labial stops /b/ and /p/). The relevance of phonetic features to speech recognition is that they have correlates in the speech signal (e.g. [2]) that play an important role in speech perception (e.g. [3]). These correlates can be reliably extracted from the signal and used in recognition [4]. In addition, the use of phonetic features as a basis for developing the acoustic parameters could lead to better understanding of contextual variability since this variability, although it may appear large at the acoustic level, involves change in one or two phonetic features (e.g. phonological rules in [5]). Such understanding may lead to building better and more economical contextual speech models than triphones, for instance.

In this paper, we consider the problem of embedding speech knowledge represented by phonetically motivated acoustic parameters in the signal representation for HMM-based recognition. Namely, we consider a set of acoustic parameters that capture the acoustic properties of the manner-of-articulation phonetic features: sonorant, syllabic, nonsyllabic, noncontinuant and fricated. These acoustic parameters were developed in earlier research [6] and fitted here to the frame-based HMM framework. Today's state-of-the-art speech recognition systems generally use a signal representation consisting of Mel-cepstral coefficients and their time derivatives. Such a signal representation contains linguistic and extralinguistic information. Consequently, in building speaker-independent speech models, statistical algorithms and large amounts of training data are relied upon to capture the phonetic message and smooth out all other information. The training data and statistical algorithms could be used more efficiently if the signal representation focuses on the phonetic information in the speech signal. Our objective in this paper is to (1) test the performance of the AP's, when used in the HMM framework (2) compare them to cepstral parameters and (3) test their robustness to speaker differences such as gender. To achieve this objective, the task of recognizing speech into the broad classes: syllabic, sonorant consonant, noncontinuant, fricative and silence was undertaken.

2. ACOUSTIC PARAMETERS

Table 1 shows the phonetic features, their corresponding acoustic correlates and acoustic parameters used in this study. The selection of the acoustic param-

This research is supported by NSF research grant # IRI-9310518

Table 1: The phonetic features, their acoustic correlates and the corresponding acoustic parameters.

Feature	Correlate	Acoustic Parameter
Sonorant	strong low-freq. energy	E0.1-0.4: 100-400 Hz eng. † E0-2_2-8: eng(0-2 KHz)-eng(2-8 KHz)
	periodic	Voicing-probability [9]
Syllabic	strong mid-freq. energy	ptd0.64-2.8: peak in 0.64-2.8 kHz eng. ptd2-3: peak in 2-3 kHz eng.
Nonsyllab.	weak mid-freq. energy	dtp0.64-2.8: dip in 0.64-2.8 kHz eng. dtp2-3: dip in 2-3 kHz eng.
Fricated	turbulence in mid to high freq. range	zcr: zero-crossing rate E0-2_2-8 R1: first autocorr. coeff. normalized by the zeroth. dtp_R1: dip-to-peak values of R1
Noncont.	Closure followed by an abrupt spectral change	Closure: E0.2-3: 0.2-3 kHz eng. †. E3-6: 3-6 kHz eng. † R1. burst : sum of positive first difference values across the STFT channels

† normalized with respect to its maximum value across the utterance.

eters was guided by the feature hierarchy of Fig. 1. Such hierarchical organization of phonetic features is advocated in modern phonology (e.g. [7]). As Fig. 1 indicates, the considered phonetic features, and consequently the corresponding parameters, allow for the recognition of the broad classes: syllabic (vowels, syllabic nasals and syllabic /l/'s), sonorant consonant (nasals and semivowels), fricative, noncontinuant (stops and affricates) and silence (silence, pauses, epenthetic silence and stop closures).

Fig. 2 depicts an example of the AP's computed from the utterance "biblical scholars" extracted from the TIMIT database. As this figure shows, the AP's capture important characteristics of the speech signal. For instance, the burst parameter in part (a) of the figure shows the highest values at the onset of the stop consonants /b/ and /k/. The voicing probability, the 0 − 2_2 − 8 kHz energy measure and the 100-400 Hz energy measure in parts (k),(l) and (m) of the figure, respectively, have their highest values during sonorant segments. Further, the peak-to-dip measures in parts (f) and (g) have their maximum values during the syllabic segments: /ih/, /el/, /aa/ and /er/. Thus, the peak-to-dip parameters coupled with those related to sonorancy will help identifying syllabic segments (c.f. Fig. 1).

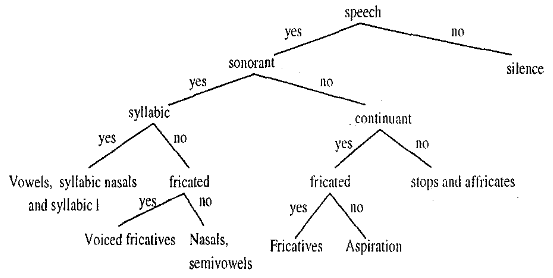

Figure 1: The hierarchy of feature organization.

The acoustic parameters for the manner phonetic-features were developed in previous work [6] based on acoustic studies and acoustic-phonetic knowledge (c.f. [8]). The philosophy adopted in defining the parameters was that they must be relative in time and/or frequency to reduce interspeaker variability. Such relative measures also take into account the relationship between different speech sounds occurring within the same utterance and spoken by the same speaker. For instance, the nonsyllabic measures are intended to measure the energy minimum in a sonorant consonant relative to the energy maximum in the preceding and/or succeeding vowel. As another example, the 100-400 Hz energy measure, being normalized with respect to the maximum in that frequency band across the utterance, accounts for the fact that the sonorant speech segments involve the voicing source of the same speaker.

3. EXPERIMENTS

In the current work, two sets of experiments were performed to evaluate the acoustic parameters and compare them to cepstral-based parameters. The task was broad-class recognition of speech into the classes: syllabic, sonorant consonant, fricative, noncontinuant and silence. For speech modeling and recognition, the HMM framework [1] was used. All experiments were conducted using the TIMIT database.

In the first set of experiments, HMM models for the broad classes were built using the TIMIT SI and SX training sentences (3573 sentences)[2]. Recognition tests were carried out using the TIMIT SI test sentences (504 sentences). In the second set of experiments, to examine robustness to interspeaker variability and more specifically to gender differences, HMM models were built using the TIMIT SI and SX training sentences spoken by females from the New England dialect region (dr1). For testing, recognition was performed on the SI and SX training sentences spoken by males from dr1. All broad-class models were context-independent

[1] implemented with HTK toolkit V1.5

[2] the remaining 123 sentences were reserved for development

3-state HMMs with diagonal-covariance Gaussian mixtures. Speech was sampled at 16 kHz and analyzed with a 5 ms frame rate for both AP's and cepstra. Energy measures in the AP's were computed from the short-time Fourier transform.

4. RESULTS AND DISCUSSION

Table 2 summarizes the experimental results where the signal representation was varied while the modeling and recognition strategy remained the same. The results for 1 and 8 mixtures show that the acoustic parameters, relative to the cepstral parameters, are better able to reduce speaker variability and target the linguistic information in the speech signal. This is deduced by comparing the small improvement in results going from 1 to 8 mixtures in the case of the AP's to the substantial improvement in case of the cepstral-based parameters. Furthermore, by comparing the AP results and the $MFCC\_E$ results to those obtained using $AP + MFCC\_E$, one can argue that the acoustic parameters contain more relevant information than the cepstral parameters. In addition, the results show that adding the first and second derivatives to the acoustic parameters improves results substantially. Preliminary analysis shows that this improved performance is due in large part to better modeling of speech dynamics, especially in the case of the stop consonants (improvement by 11% with 8 mixtures).

We expect the performance of the HMM system with acoustic parameters to improve further once other parameters including those related to place of articulation are added. Preliminary analysis suggests that this additional information will aid recognition in two ways. First, adjacent sounds that share the same manner of articulation can be distinguished based on place information. At present, two adjacent fricatives such as /sš/ are often recognized as one long fricative. In scoring, since timing information is not fully accounted for, one of the fricatives is considered deleted. Second, sounds whose manner-of-articulation features change from their canonical form due to context, may still be correctly recognized if other features are not altered. For example, it is often the case that voiced obstruents such as the weak voiced fricative /v/ are realized as sonorants when they occur in an intervocalic position [10]. Presently, the HMM system with acoustic parameters is constrained to recognize such a /v/ as a sonorant consonant. Although such a classification is correct, it is scored as incorrect because sonorant /v/'s are not distinguished from the fricated /v/'s in the TIMIT labels and they are all considered as fricatives in scoring. In any case, when a full feature-based representation is used so that phone recognition is performed, it is expected that /v/ will be correctly recognized despite its change in manner.

Table 2: Recognition results. $MFCC\_E$ refers to 12 Mel-cepstral coefficients & log energy, $MFCC\_E\_\delta 1\_\delta 2$ refers to $MFCC\_E$ & their 1st and 2nd derivatives, AP refers to acoustic parameters, $AP\_\delta 1\_\delta 2$ refers to AP and their 1st and 2nd derivatives. Each entry contains % correct/% accuracy.

Signal Representation	1 mix	8 mix
$MFCC\_E$	68.2/61.8	73.3/65.2
$MFCC\_E\_\delta 1\_\delta 2$	73.5/63.7	82.8/71.5
AP	75.2/63.9	77.5/66.3
$AP\_\delta 1\_\delta 2$	78.5/68.1	84.1/71.5
$AP + MFCC\_E$	75.2/63.9	78.1/66.7

Table 3: Recognition results using 8 mixtures. Training done with speech produced by females. Recognition done with speech produced by males.

Signal Representation	%correct/%accurate
$MFCC\_E\_\delta 1\_\delta 2$	81.13/66.75
$AP\_\delta 1\_\delta 2$	83.3/70.7

Table 3 summarizes the experimental results obtained when the recognizers were trained on speech produced by females and tested on speech produced by males. Compared to the results obtained with the cepstral parameters, the results obtained with the acoustic parameters are much closer to the corresponding results listed in Table 2, indicating more robustness to gender variability. Additional experiments were conducted with the AP's and their derivatives. In the first experiment, a 1 mixture system, instead of 8 mixtures, was trained using all SI and SX female sentences in the TIMIT training set and tested on all SI male sentences in the TIMIT test set. A 1% performance degradation was observed compared to the system trained on both males and females. In the second experiment, the same system was trained on all SI and SX male sentences in the TIMIT training set and tested on the SI female sentences in the test set. No degradation in performance was observed. These experiments indicate the robustness of the AP's to gender differences.

5. CONCLUSION

The recognition results show that, compared to cepstral-based parameters, the phonetically-based acoustic parameters are better able to target linguistic information. This ability to emphasize phonetically-relevant information while discarding the extra-linguistic speech properties is highlighted by the better results obtained from the gender experiment and by comparing the results obtained using 1 mixture to those obtained with 8 mixtures.

In this research, we considered manner-of-articulation phonetic features and related acoustic parameters. Currently, we are studying the acoustic properties related to place of articulation and developing parameters that capture these properties. Once these parameters are developed, phoneme recognition experiments will be conducted. In addition, methods that optimize the selection of acoustic parameters based on objective criteria, as opposed to our currently adopted histogram analysis method, are being examined. Further, as it may be clear from this paper, the number of acoustic parameters needed for phoneme recognition will increase the dimensionality of the signal representation in comparison to the cepstral-based representation. Thus, reducing the parameters to the phonetic feature dimensions they represent will be considered as a way of dealing with the dimensionality increase if it proves to be a problem.

6. REFERENCES

[1] Chomsky, Noam and Halle, Morris, "The Sound Pattern of English," The MIT Press, Cambridge, Massachusetts, U.S.A, 1968.

[2] Stevens, K., "Acoustic Correlates of Some Phonetic Categories, *JASA,* 1980, 68, 836-842.

[3] Hedrick, Mark S. and Ohde, Ralph N., "Effect Of Relative Amplitude Of Frication On Perception Of Place Of Articulation," *JASA,* Vol. 94, No. 4, Oct. 1993.

[4] Espy-Wilson, C., "A Feature-Based Approach to Speech Recognition,"*JASA,* 1994, Vol. 96, 65-72.

[5] Oshika, B.T., Zue, V. W., Weeks, R.V, Neu, H. and Aurbach, J., " The Role of Phonological Rules in Speech Understanding Research," *IEEE trans. ASSP* Vol. ASSP-23, No. 1, Feb. 1975.

[6] Bitar, N. and Espy-Wilson, C., "A Signal Representation of Speech Based on Phonetic Features," *Proc. of 1995 IEEE Dual-Use Tech. and Appl. Conf.,* SUNY Tech., Utica/Rome.

[7] Clements, N., "The Geometry of Phonological Features," *Phonol. Yearbook 2,* 1985, 225-252.

[8] Zue, V., "Spectrogram Reading Notes," MIT.

[9] *ESPS 5.1,* Entropic Research Laboratory Inc.

[10] Olive, Joseph et. al., *Acoustics of American English Speech,* Springer-Verlag, New York, 1993.

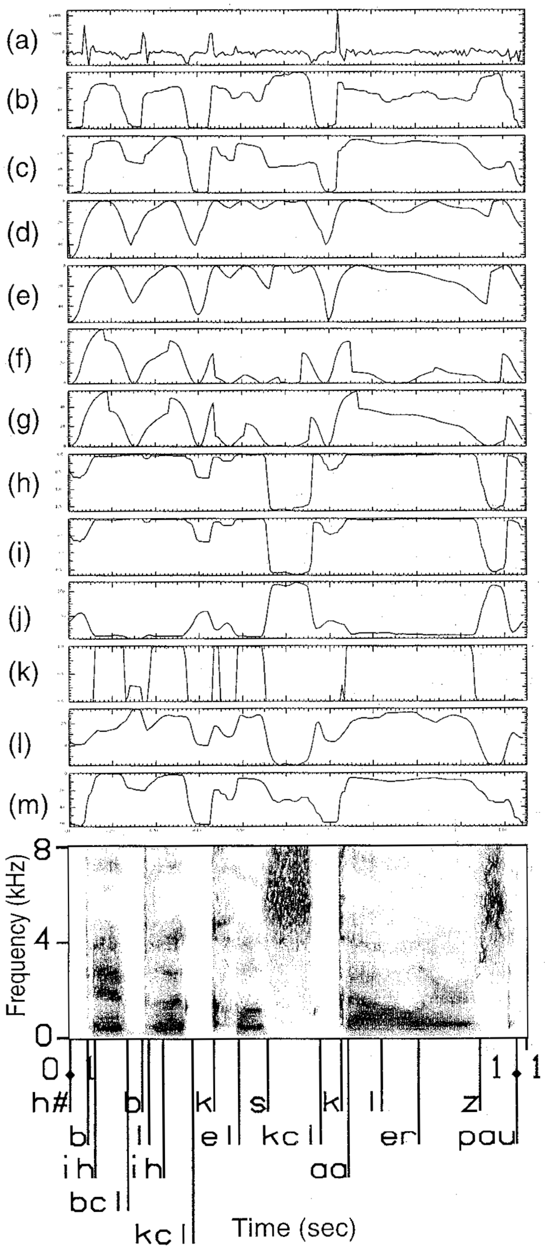

Figure 2: This figure illustrates the set of parameters listed in Table 1. These parameters are: (a) burst, (b) E3-6, (c) E0.2-3, (d) ptd_0.64-2.8, (e) dtp_0.64-2.8, (f) ptd_2-3, (g) dtp_2-3, (h) dtp_R1, (i) R1, (j) zcr, (k) voicing-probability, (l) E0-2_2-8 and (m) E0.1-0.4.