An Event-Based Acoustic-Phonetic Approach for Speech Segmentation and E-Set Recognition

Amit Juneja, Om Deshmukh and Carol Espy-Wilson
Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742
http://www.ece.umd.edu/~juneja

1. INTRODUCTION

The E-set utterances - B, C, D, E, G, P, T, V and Z - form a set of highly confusable sounds. Accurate recognition of these sounds is the most significant step for the improvement of recognition performance on the connected 'alphadigit' task that includes sequences of spoken digits and letters. We extended our earlier event-based recognition system (EBS) [1] and applied it to this difficult recognition task. In particular, we focused on the ability of EBS to distinguish among the stop consonants in the alphabets B, D, P, and T and among the fricatives in C, Z and V and the affricate in G.

2. EBS

In EBS, the speech signal is first segmented into the broad classes: vowel, sonorant consonant, strong fricative, weak fricative and stop. This segmentation is based on acoustic events (or landmarks) obtained in the extraction of acoustic parameters (APs) associated with the manner phonetic features *sonorant, syllabic, continuant and strident*, in addition to silence. (Results on the performance of the broad class recognizer are given in [1]). The manner acoustic events are then used to extract parameters relevant for the *voice* phonetic feature and for the place phonetic features. In particular, APs for the place features *labial* and *alveolar* are extracted for stops; and APs for the place feature *anterior* are extracted for strident fricatives. This phonetic feature hierarchy, as advocated in [2], is shown in Figure 1 for the special case of the E-set.

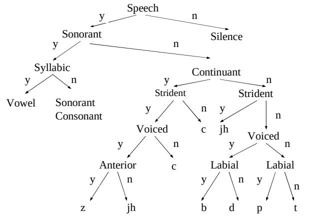


Figure 1: Phonetic feature hierarchy tailored for the E-set

3. DATABASE

The E-set utterances (B,C,D,E,G,P,T,V,Z) from the TI46[3] database [2] were used for this project. To develop EBS, the TI46 training set was used which consists of these words spoken by 16 speakers, 8 males and 8 females. For testing, the TI46 test set was used. It consists of a set of different utterances from the same speakers.

4. APs

Table 1 shows the acoustic parameters designed to extract the acoustic correlates for the phonetic features needed to recognize voicing and place for the strident fricatives and stops. Note that Ahi-A23 and Av-Ahi are measures similar to those proposed by Stevens [4]. Ahi-A23 measures the spectral tilt and Av-Ahi measures the spectral prominence of F1 (first formant) relative to the high frequency peak of the consonant. We have modified these parameters so that they depend on the average of the third formant (F3) over the utterance to achieve vocal tract normalization.

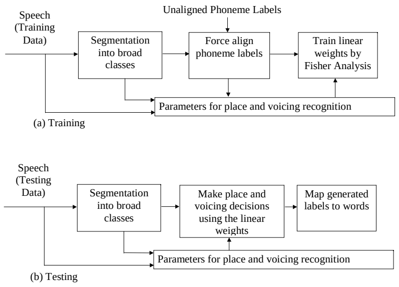
Manner	Phonetic Feature	Parameter
Stops	Voicing	Voicing onset time (VOT), probability of voicing [5], zero crossing rate, F2 and F3 transitions
	Consonantal	Strength of consonant onset
	Labial/Alveolar	Ahi-A23, Av-Ahi
Fricative	Voicing	Duration, probability of voicing
	Strident	Energies in three equal frequency bands between 2000Hz and end of the spectrum, zero crossing rate, total energy
	Alveolar	Ahi-A23, Energy in the band [F3-187 Hz, F3+781 Hz]

Table 1: Parameters extracted by EBS for place and voicing recognition of stops and fricatives

5. ACOUSTIC PARAMETERS AND EBS

5.1 Training and Testing

EBS uses a fuzzy logic-based approach to explicitly segment the speech into broad classes [5]. To find the optimal linear weights for combining the APs for place and voicing decisions, we generated an automatic transcription of the E-set utterances in TI46 training data. To do so, phoneme labels were formed for the E-set utterances and they were forced aligned by comparison with the broad class labels generated by EBS.


Speech (Training Data)
Unaligned Phoneme Labels
Segmentation into broad classes
Force align phoneme labels
Train linear weights by Fisher Analysis
Parameters for place and voicing recognition
(a) Training

Speech (Testing Data)
Segmentation into broad classes
Make place and voicing decisions using the linear weights
Map generated labels to words
Parameters for place and voicing recognition
(b) Testing

Figure 2: Training and Tesrting in EBS

At this stage, Fischer linear discriminant analysis (LDA) was applied to train the linear weights on the parameters for the place and voicing recognition of stops and fricatives. Although Ahi-A23 and Av-Ahi are the most discriminative cues for the recognition of alveolar and labial place phonetic features, Figure 3 shows how a combination of these APs with others listed in Table 1 improves the separation of the phonemes /b/ and /d/. In Figure 3(a), the instances of /b/ and /d/ are projected from (Ahi-A23, Av-Ahi) space to one-dimensional space using LDA. In Figure 3(b), the instances of /b/ and /d/ are projected from the higher dimensional space, that includes other APs from Table 1, to one-dimensional space using LDA. Figure 2 shows the training and testing schemes.

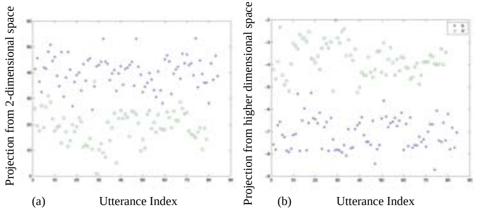


*Figure 3: Combination of the parameters Ahi-A23 and Av-Ahi along with other acoustic-phonetic parameters using LDA increases the separation of the phoneme segments /b/ and /d/. Legends: □ is for /d/ and * is for /b/.*

5.2 Results

Table 2 shows the accuracy of EBS in different stages of classification. The system gives high accuracy for place and voicing recognition, and we are currently working on the improvement of the broad class segmentation system. EBS gives 75.7% E-set word accuracy on TI46 test data. Previously obtained best results [5] on the E-set using context-independent hidden Markov models (HMMs) were at 75.6%.

Phoneme	Manner Accuracy	Voicing Accuracy	Place Accuracy
/b/	82.72	100	94.87
/c/	98.43	94.33	98.76
/d/	82.93	97.13	97.60
/Q/	53.56	94.38	-
/jh/	85.83	100	89.61
/p/	92.91	87.87	94.06
/t/	85.88	100	99.54
/v/	67.71	100	-
/z/	92.10	96.44	100

Table 2: Performance of EBS at different classification stages. All results in percentages

6. APs and HMMs

In developing EBS, we came up with highly discriminative measures that can be used as speaker-independent APs for HMM-based speech recognition systems [7]. Results of HMM-based recognition using acoustic phonetic parameters on the phoneme recognition of the E-set are shown in Table 6..

Front-end	Number of parameters	Phoneme accuracy (%)
MFCC	39	82.45
Acoustic-phonetic	23	85.14

Table 6. Comparison of MFCC front-end with acoustic-phonetic front end

7. CONCLUSION

We have proposed a landmark-based method for speech segmentation and recognition. The method utilizes acoustic-phonetic knowledge to find measures that have higher discriminative capacity for speech recognition than cepstral measures.

8.REFERENCES

[1] Bitar, N. and Espy-Wilson, C., "A signal representation of speech based on phonetic features.", 5th Dual-Use Technologies and Applications Conference, May 22-25, 1995.

[2] Halle & Clements, Problem Book in Phonology: A Workbook for Introductory Courses in Linguistics and Modern Phonology

[3] Mark Liberman, TI46-Word, http://www.ldc.upenn.edu/

[4] Stevens, K.N., and S.Y. Manuel, "Revisiting Place of Articulation Measures for Stop Consonants: Implications for Models of Consonant Production", In the Proceedings of XIV International congress of phonetic sciences Vol.2 pp 1117-1120.

[5] Philipos, L. and Spanias, A., "High-Performance Alphabet Recognition", IEEE Transactions Speech and Audio Processing, Vol. 4, no. 6, pp. 430-445, November 1996.

[6] Deshmukh, O., Espy-Wilson, C. and Juneja, A, "Acoustic-Phonetic Speech Parameters for speaker independent specch recognition", ICASSP 2002, SP-P08.06

ACKNOWLEDGEMENTS

This work was supported by NSF grant #SBR-9729688