# An Event-Based Acoustic-Phonetic Approach For Speech Segmentation And E-Set Recognition

**Amit Juneja and Carol Espy-Wilson**
Department of Electrical and Computer Engineering, University of Maryland,
College Park, MD 20742, USA
Email: juneja@glue.umd.edu, espy@glue.umd.edu

## ABSTRACT

In this paper, we discuss an automatic event-based recognition system (EBS) that is based on phonetic feature theory and acoustic phonetics. First, acoustic events related to the manner phonetic features are extracted from the speech signal. Second, based on the manner acoustic events, information related to the place phonetic features and voicing are extracted. Most recently, we focused on place and voicing information needed to distinguish among the obstruents /b/, /d/, /p/ /t/, /s/, /z/ and /jh/. Using the E-set utterances from the TI46 test database, EBS achieved place accuracy of 96.35% and voicing accuracy of 96.53% for the above mentioned obstruents. The recognition accuracy of EBS is presented at its different stages of classification. The applications of the system apart from phoneme recognition are also discussed.

## 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 utterances B, D, P, and T.

## 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 [1,13]).

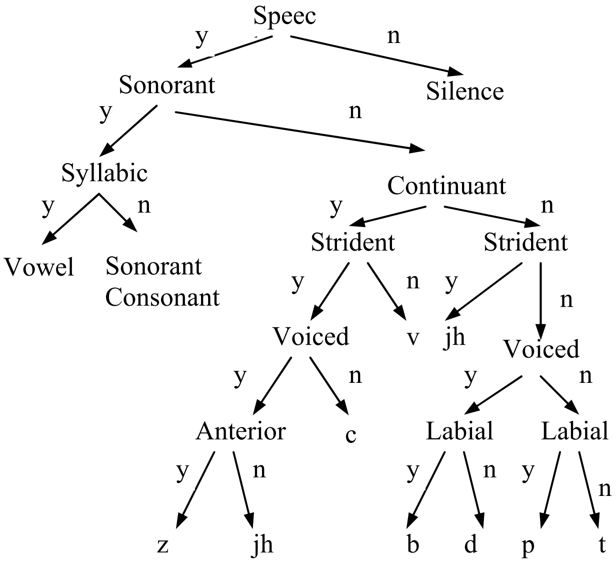


*Figure 1: Phonetic feature hierarchy tailored for the E-set. 'y' = yes and 'n' = no*

The manner acoustic events are then used to extract APs relevant for the voiced phonetic features 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. Note that this strategy is very different from the hidden Markov model (HMM) framework where all parameters are computed in every frame. Instead, EBS uses the manner landmarks to determine (1) which APs are computed for place and voicing and (2) the time region used to extract this information.

## 3. DATABASE

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

## 4. ACOUSTIC PARAMETERS

Table 1 shows the APs 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 modifed these parameters with respect to the average frequency 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 [6] |
| | Strident | Energies in three equal frequency bands between 2000Hz and end of the spectrum, zero crossing rate, total energy |
| | Anterior | Ahi-A23, Energy in the band [F3-187 Hz, F3+781 Hz] |

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

## 5. ACOUSTIC PARAMETERS AND EBS

### 5.1 Training

EBS uses a fuzzy logic-based approach to explicitly segment the speech into broad classes [1,13]. To find the optimal linear weights for combining the APs, 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 force-aligned by comparison with the broad class labels generated by EBS. For example, if EBS produces the sequence of labels (STOP, VOWEL) for the utterance T, then the phoneme labels /t/ and /iy/ are given the exact locations of the STOP and VOWEL respectively. At this stage, Fischer linear discriminant analysis (LDA) [12] was applied to train the linear weights on the parameters for the place and voicing recognition of stops and fricatives. Note that this approach can be extended to make distinctions between different vowels and sonorant consonants (nasals, liquids, glides, etc.). The training procedure is shown in Figure 2(a).

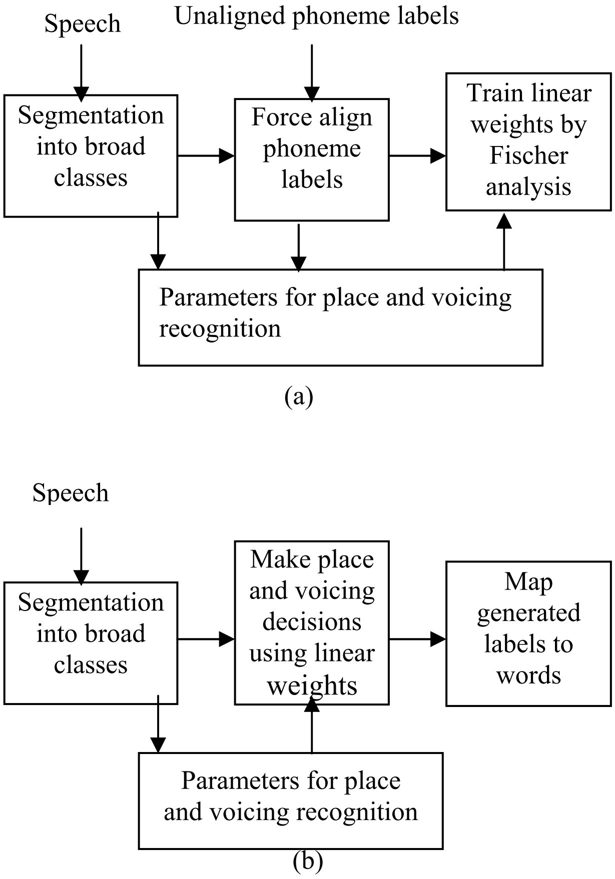
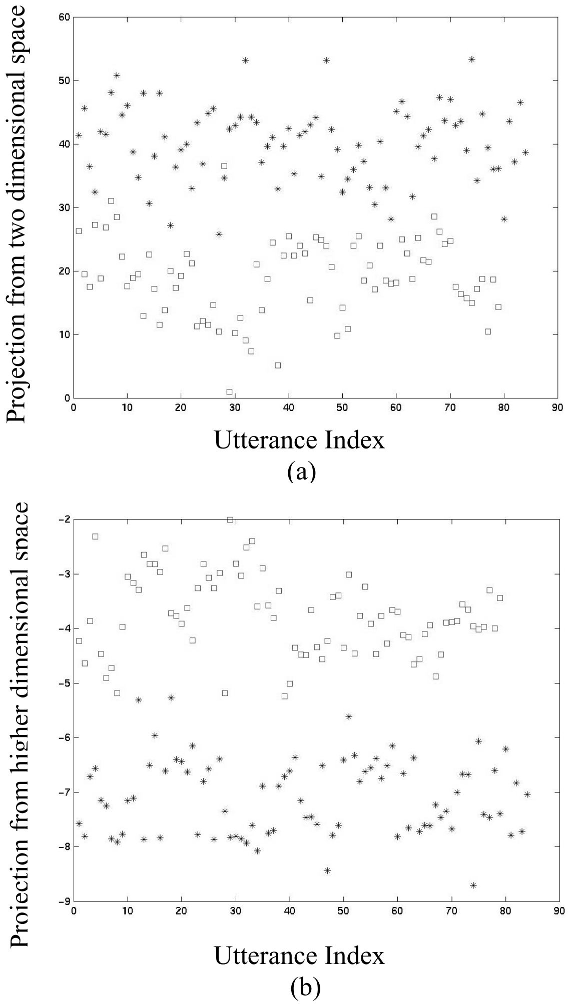

Speech
Unaligned phoneme labels
Segmentation into broad classes
Force align phoneme labels
Train linear weights by Fischer analysis
Parameters for place and voicing recognition
(a)
Speech
Segmentation into broad classes
Make place and voicing decisions using linear weights
Map generated labels to words
Parameters for place and voicing recognition
(b)

*Figure 2 Training (a) and testing (b) schemes*

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. This results in an increased separation of these voiced stops.

### 5.2 Testing and Scoring

The broad classification and trained linear weights are used in testing, as outlined in Figure 2(b). The generated output labels are now mapped to E-set words - B, C, D, E, G, P, T, V, Z – by merging labels. For example (Figure 4)



(a)

(b)

*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 /b/ and * is for /d/.*

voice bars (VBs) are often detected in the closure regions of voiced stops, the /y/ off-glide of the vowel /iy/ is detected as sonorant consonant (SC) and weak frication (WF) noise occurs at the end of the vowel /iy/. This fine detail is often useful in disambiguating similar sounds. For example, the stop burst in Figure 4 is detected as /t/, apparently because of its long VOT. But /t/ cannot follow a voice bar, therefore, it is replaced by /d/. In addition, the labels ‘SC’ and ‘WF’ are merged with V. So the scoring program decides that the utterance is D.

### 5.3 Results

Table 2 shows the accuracy of EBS in different stages of classification. The system gives high accuracy for place and voicing recognition. We are currently working on the improvement of the broad class segmentation system.

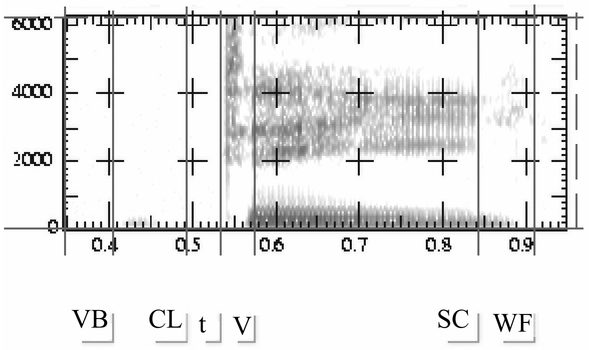


*Figure 4: Spectrogram and the labels generated by EBS on utterance D. VB = voicebar, CL = closure, V = vowel, SC = sonorant consonant, WF = weak fricative*

Especially, we are concentrating on better manner recognition of weak fricatives such as /v/ and /f/. The manner recognition of /v/ is hard because it manifests itself in a number of different ways, as illustrated in [10]. Glottal stops, represented by /Q/, are currently distinguished from voiced stops - /b/ and /d/ - only by the lack of formant transitions following the stop burst. The place accuracy of glottal stop /Q/ is at 64.22% because the formant tracker sometimes fails in the transition regions. EBS gives 75.7% E-set word accuracy on TI46 test data.

| Phoneme | Manner Accuracy | Voicing Accuracy | Place Accuracy |
|---|---|---|---|
| /b/ | 82.72 | 100 | 94.87 |
| /s/ | 98.43 | 94.33 | 98.76 |
| /d/ | 82.93 | 97.13 | 97.60 |
| /Q/ | 84.00 | 94.38 | 64.22 |
| /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 are in percentages.*

## 6. OTHER APPLICATIONS

The explicit segmentation capability of EBS has been used for enhancement of impaired speech [6]. Furthermore, in developing EBS, we came up with highly discriminative measures that can be used as speaker independent acoustic parameters for HMM based speech recognition systems [7,8]. We are working on using EBS for automatic phoneme alignment in databases because the locations of labels generated by EBS match closely to the handtranscribed labels [10]. Lastly, segmentation of sonorant regions into vowels, sonorant consonants and nasals can help in better tracking of formants [9].

## 7. CONCLUSION AND FUTURE WORK

We have proposed a landmark-based method for speech segmentation and recognition. The method utilizes acoustic-phonetic knowledge to find measures that have high discriminative capacity for phoneme recognition. We obtain a high recognition accuracy of place and voicing of the obstruents - /b/, /d/, /p/, /t/, /s/, /z/ and /jh/. In the future, we plan to improve the EBS broad class segmentation as well as incorporate measures for more place features for stops and fricatives. Broad class segmentation, especially the detection of stops and weak fricatives, may be improved further by using temporal parameters [11]. We plan to replace the ESPS formant tracker [5] by our new formant tracker [9] that is based on acoustic phonetics knowledge and dynamic programming. A better formant tracker may help in the improvement of place recognition of glottal stops. We will also study coarticulation and incorporate that knowledge into EBS. A statistical and probabilistic framework is being developed for EBS to enable the system to carry out large recognition tasks [13,14].

## 8.REFERENCES

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

[2] Halle, M., and Clements, G. N., ”Problem book in phonology. Cambridge”, MA: MIT Press, 1983.

[3] http://www.ldc.upenn.edu/Catalog/LDC93S9.html

[4] Stevens, K.N., and Manuel, S. Y.,“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] ESPS (Entropic Signal Processing System 5.3.1), Entropic Research Laboratory, http://www.entropic.com

[6] Espy-Wilson, C., Chari, V., MacAuslan, J., Huang, C., and Walsh, M.,“Enhancement of Electrolaryngeal Speech by Adaptive Filtering.” Journal of Speech, Language, and Hearing Research, 41, 1253-64

[7] Deshmukh, O., Espy-Wilson, C. and Juneja, A., “Acoustic–phonetic speech parameters for speaker-independent speech recognition”, ICASSP2002, May 13-17, 2002, Orlando Florida.

[8] Bitar, N. and Espy-Wilson, C., “Knowledge-based parameters for HMM speech recognition”, Conference Proceedings, ICASSP 1996, Volume 1, I29.

[9] Xia, K. and Espy-Wilson, C., “A New Formant Tracking Algorithm based on Dynamic Programming”, Proceedings of International Conference on Spoken Language Processing, Oct. 2000

[10] Illustrations. http://www.ece.umd.edu/~juneja

[11] Saloman, A., and Espy-Wilson, C., “Automatic Detection of Manner Events for a Knowledge-Based Speech Signal Representation”, Proc.of Eurospeech, Sept. 1999, Budapest Hungary**.**

[12] Duda, R. O., Hart. P., E., and Stork, D. G.,”Pattern Classification” (2nd ed), Wiley, New York, NY 2000.

[13] Juneja, A. and Espy-Wilson, C., "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning.", in the proceedings of International Conference on Neural Information Processing, 2002, Volume 2, Page 726-730

[14] Juneja, A., and Espy-Wilson, C., "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines", submitted to International Joint Conference on Neural Networks, Portland, Oregan, 2003.

**ACKNOWLEDGMENTS**
This work was supported by NSF grant #SBR-9729688