Acoustic-phonetic Approach to Speech Recognition Based on Landmark Detection

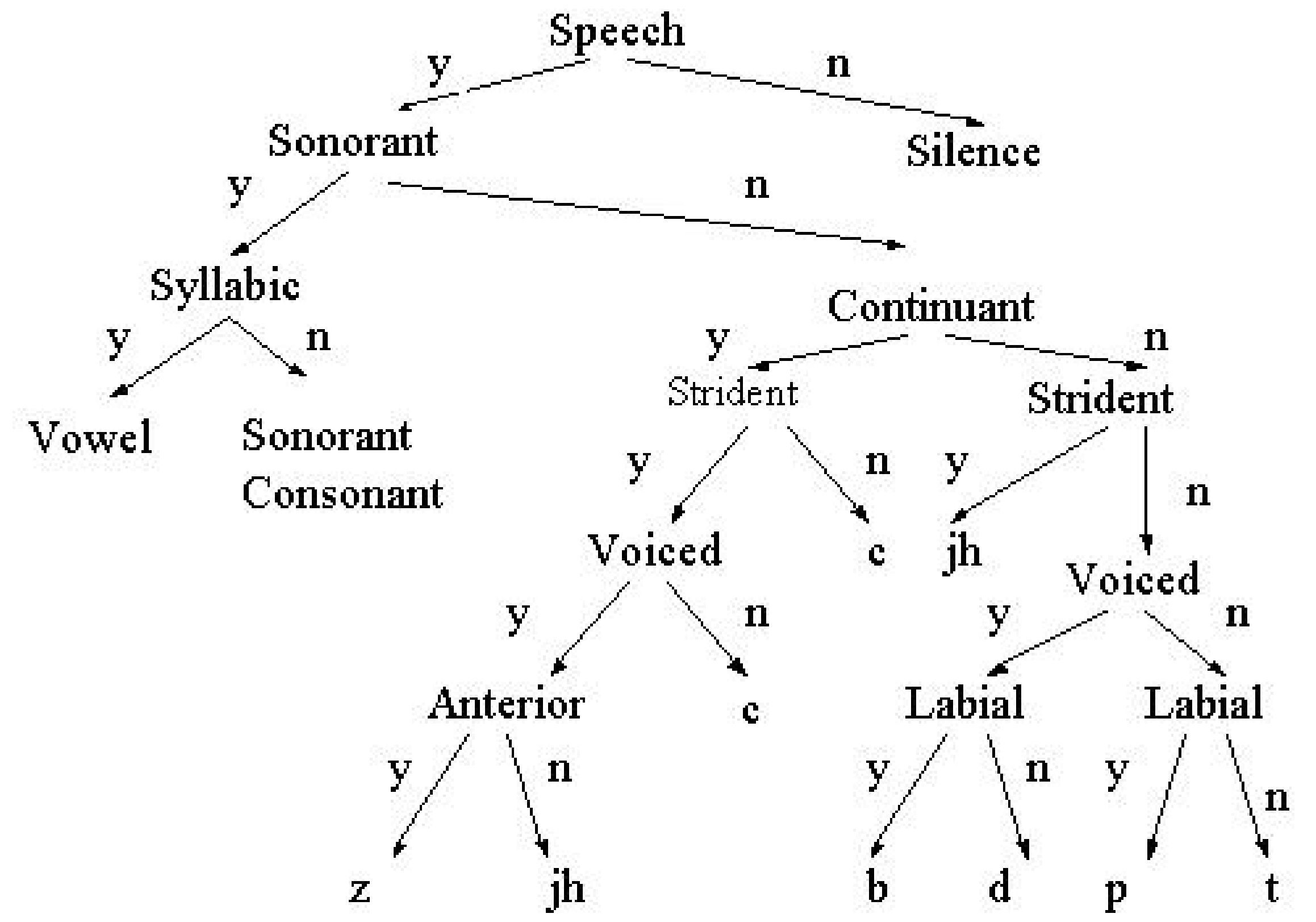
Amit Juneja and Carol Espy-Wilson

1. Introduction

We discuss an event-based recognition system (EBS) which combines phonetic feature theory, acoustic phonetics and statistical processing. Unlike the state-of-the-art HMM system, EBS is database independent and speaker independent. Thus, it does not have to be retrained. In addition, EBS does not perform frame-based recognition. Instead, recognition occurs around linguistically relevant landmarks.

2. Method

The phonetic feature hierarchy [1] shown in Fig. 1 shows the recognition strategy used in EBS. First, acoustic parameters (APs) related to the manner phonetic features *sonorant, syllabic, continuant* and *strident* are used to segment the speech signal into the broad classes: **vowel, sonorant consonant, strong fricative, weak fricative** and **stop**. Second, landmarks obtained from the manner recognition guide the extraction of APs related to the voice phonetic features and the place 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.



Speech
y
n
Sonorant
Silence
y
n
Syllabic
Continuant
y
n
y
n
Vowel
Sonorant Consonant
Strident
Strident
y
n
y
n
Voiced
c
jh
Voiced
y
n
y
n
Anterior
c
Labial
Labial
y
n
y
n
y
n
z
jh
b
d
p
t

Figure 1: Phonetic feature hierarchy tailored for the E-set utterances – B,C,D,E,G,P,T,V,Z. 'y' = yes and 'n' = no

The APs are based on exact measures taken from the time-frequency representation of the speech waveform. Only relative measures (as opposed to absolute measures) are used to minimize speaker differences. Knowledge-based rules and linear discriminant analysis are used to determine how the APs are combined. Figure 2 shows the output of EBS and some of the intermediate waveforms for three sample utterances.

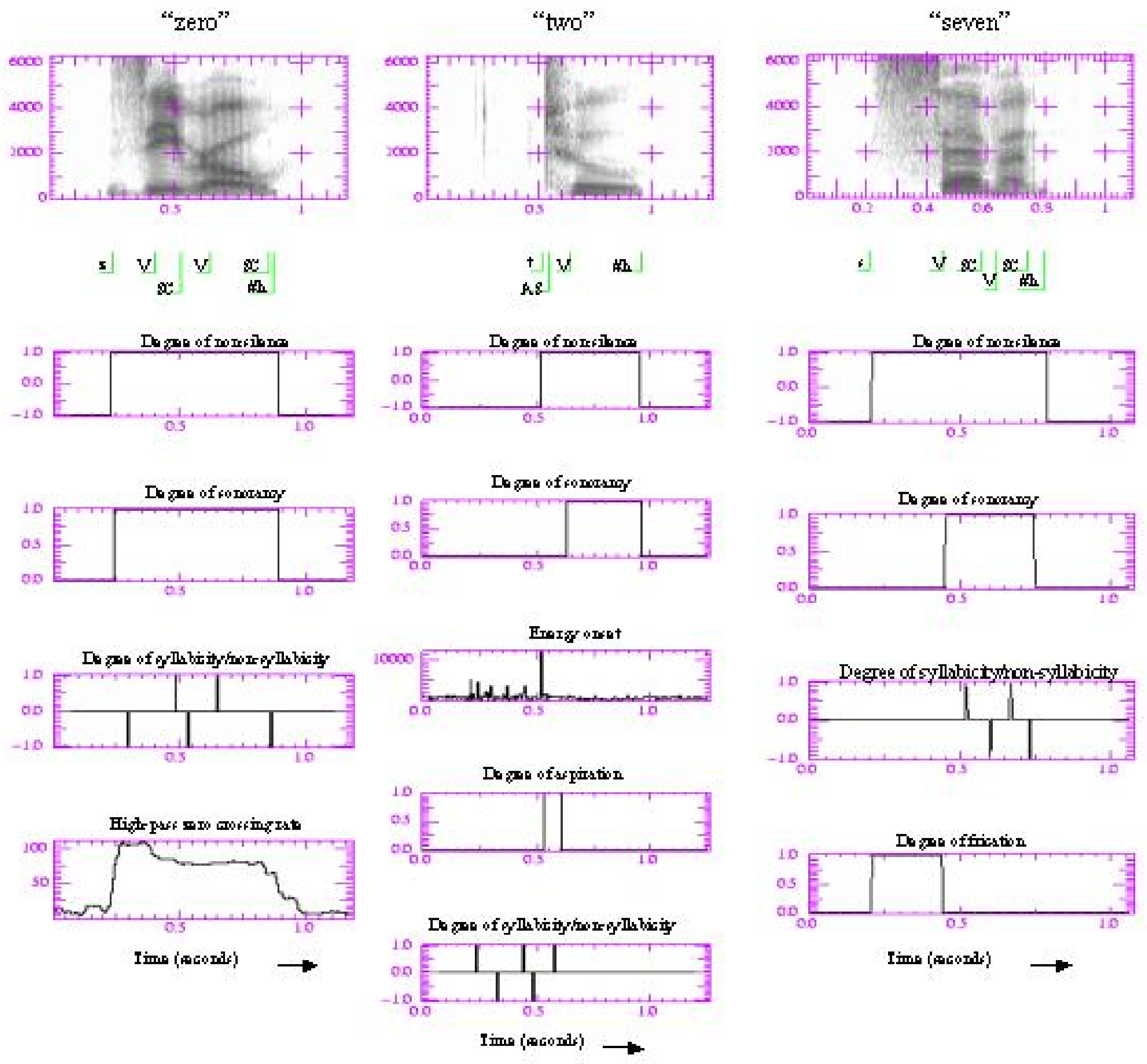


Figure 2. Output of EBS for three utterances.

3. Results

EBS was tested with the E-set (B,C,D,E,G,P,T,V,Z) utterances taken from the TI-46 test corpus. Overall, EBS obtained a 75.7% word accuracy which is slightly better than the best context-independent HMM recognition system [2]. Table 1 shows performance results at different stages of classification. As can be seen, EBS gives high accuracy for place and voicing recognition. However, manner recognition needs to be improved. This problem is due to variability, particularly during weak fricatives and weak stops. At present, we are exploring the use of Support Vector Machines (SVMs) [3] to better capture and model this variability.

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/	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 1: Performance of EBS at different classification stages. All results in percentages

4. Future Work

Table 1 shows that EBS does not do well in the manner recognition of weak fricatives (particularly the phoneme /v/). As Fig. 3 shows, the surface manifestation of this sound can vary considerably. At present, we are planning to deal with such extensive variability by combining acoustic-phonetic knowledge with the non-linear decision capabilities of Support Vector Machines.

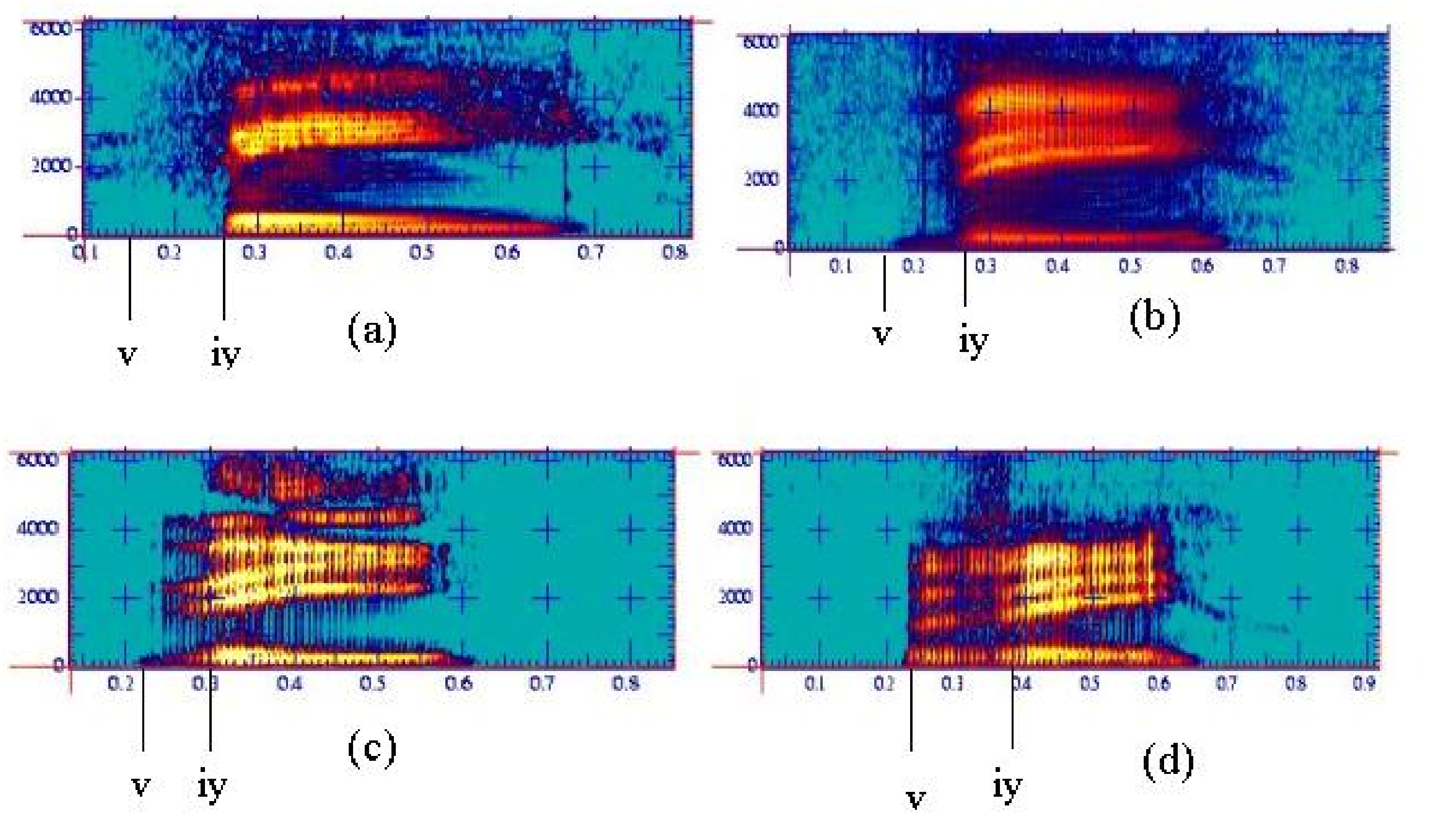


Figure 3: Different manifestations of the phoneme 'v'. (a) frication only. (b) transient, weak frication and voice bar. (c) sonorant. (d) sonorant and frication.

5. References

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

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

[3] Kecman, V., "Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models", The MIT Press, 2001

Acknowledgements

This work was supported by NSF grant #SBR-9729688