

Speech Segmentation Using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines

Amit Juneja and Carol Espy-Wilson

Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742, USA
juneja@glue.umd.edu, espy@glue.umd.edu

Abstract— We propose a method that combines a probabilistic phonetic feature hierarchy with support vector machines for segmentation of continuous speech into five classes - vowel, sonorant consonant, fricative, stop and silence. We show that by using the hierarchy, only four binary classifiers are required to recognize the five classes. Due to the probabilistic nature of the hierarchy, the method overcomes the disadvantage of the traditional acoustic-phonetic methods where the error is carried down the hierarchy. In addition, the hierarchical approach allows the use of comparable amount of training data of two classes that each binary classifier is designed to discriminate. The segmentation method with 13 knowledge based parameters performs considerably better than a context-independent Hidden Markov Model (HMM) based approach that uses 39 mel-cepstrum based parameters.

I. INTRODUCTION

Support Vector Machines (SVMs) [1] have been shown to be very effective in feature detection [2], [3] and classification of segmented phonemes [4], [5] in continuous speech. SVMs have attractive properties for pattern classification, for example, capability of learning from small amount of data, capacity control, and elegant handling of high dimensional data. But the use of SVMs or neural networks (NNs) as a unified statistical framework in automatic speech recognition remains limited because of inferior modeling of time-varying dynamics and coarticulation. In an acoustic-phonetic approach to speech recognition, binary decision making is involved at certain linguistically motivated landmarks and a classifier for variable length sequences of observation vectors is not required. We proposed a segmentation method for continuous speech using acoustic-phonetic knowledge and five separate binary SVM classifiers [6]. In this paper we introduce the concept of a probabilistic phonetic feature hierarchy and incorporate it in the recognition task.

In our event-based system (EBS), speech is first segmented into several broad classes - vowels, sonorant consonants (nasals and semi-vowels), fricatives, stops and silence using the phonetic feature hierarchy shown in Figure 1. Next, the parameters for place and voicing are extracted to decide upon the phonemes. We show that for segmentation into five broad classes we do not need five classifiers because the hierarchy in Figure 1 can be exploited in a probabilistic manner. The idea can be extended to phoneme recognition because all phonemes can be represented by the presence or absence of 20 odd phonetic features [7]. EBS is a bottom-up speech recognizer

because it first explicitly carries out the recognition at the level of phones, which is more like human speech recognition [8]. In contrast, Hidden Markov Model (HMM) based state-of-the-art speech recognizers are top-down [9], [10].

We have shown before [11] that knowledge based acoustic parameters (APs) are more speaker independent and give comparable performance for digit recognition task than the mel-cepstrum based coefficients. We have used 13 of those knowledge-based APs that are acoustic correlates of the manner phonetic features - sonorant, syllabic, continuant, in addition to silence - for the segmentation task. Classifiers for the different phonetic features use APs that are correlates of the corresponding phonetic features. This method has three added advantages - (1) Not all APs are used for all decisions, (2) Since the APs have a strong physical interpretation, it is easy to pinpoint the source of error in such a recognition system. That is, it is easy to tell whether the pattern matcher has failed, the knowledge based APs need to be refined, or if there is some source of variability that we haven't accounted for. (3) The method can easily take advantage of years of research that has gone into acoustic phonetics as well as signal processing based on human auditory models, for example, [12].

II. DATABASE

The TIMIT database [13] was used for the experiments presented in this paper. The phonetically rich 'si' and 'sx' sentences of all dialect regions from the training set were used for training and the 'si' sentences of all dialect regions from the test set were used for testing.

III. SUPPORT VECTOR MACHINES

SVMs are learning machines for pattern classification and regression tasks based on statistical learning theory [1]. Given a set of training vectors $\{\mathbf{x}_i\}_{i=1}^l$, and the corresponding class labels $\{y_i\}_{i=1}^l$ such that

$$y_i \in \{-1, +1\} \text{ and } \mathbf{x}_i \in \mathbb{R}^n,$$

SVMs select a set of support vectors $\{\mathbf{x}_i^{SV}\}_{i=i}^{N_{SV}}$ that is a subset of the training set $\{\mathbf{x}_i\}_{i=1}^l$ and find an optimal decision function

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N_{SV}} y_i \alpha_i K(\mathbf{x}_i^{SV}, \mathbf{x}) - b\right)$$

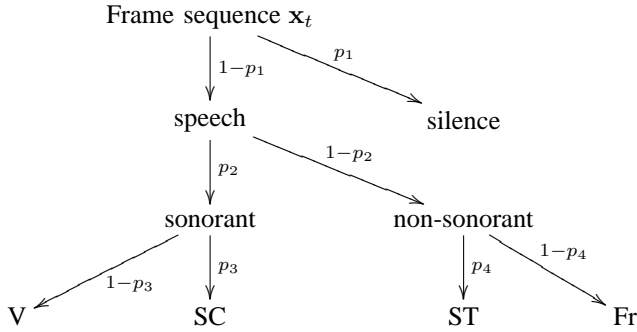


Fig. 1. Probabilistic phonetic feature hierarchy

where K is an a priori chosen kernel function. The weights α_i , the set of support vectors $\{\mathbf{x}_i^{SV}\}_{i=1}^{N_{SV}}$ and the bias term b are found from the training data using quadratic optimization methods. Many methods have been suggested to convert SVM outputs to probabilities, but in our project we currently clip the output between -1 and +1 and map the result to [0,1]. We have used radial base function (RBF) kernels and linear kernels in these experiments. For RBF kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma|\mathbf{x}_i - \mathbf{x}|^2)$$

where the parameter γ is chosen empirically by cross-validation from the training data. For the linear kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i \cdot \mathbf{x} + 1$$

The experiments in this project were carried out using the SVM Light toolkit [14], which provides very fast training of SVMs.

IV. METHOD

Figure 1 shows the phonetic feature hierarchy suggested in [7]. We have made the hierarchy probabilistic by assigning probabilities to each branch. The hierarchy is shown only to the level of five broad classes - vowel (V), sonorant consonant (SC), fricative (Fr), stop (ST) and silence (SIL). Consider a sequence of parameter vectors

$$\mathbf{x}_t = \{\mathbf{o}_{t-s}, \mathbf{o}_{t-s+1}, \dots, \mathbf{o}_{t+e}\}$$

at a given instant t where s previous frames and e following frames are used along with the current frame \mathbf{o}_t for analysis. Assume that the frame at time t lies in the region of one of the broad classes. We can write the posterior probability of the frame being part of a vowel at time t as

$$\begin{aligned} P(V|\mathbf{x}_t) &= P(\text{speech}, \text{sonorant}, \bar{SC}|\mathbf{x}_t) \\ &= P(\text{speech}|\mathbf{x}_t)P(\text{sonorant}|\text{speech}, \mathbf{x}_t) \\ &\quad (1 - P(SC|\text{sonorant}, \mathbf{x}_t)) \\ &= (1 - p_1)p_2(1 - p_3) \end{aligned}$$

and similarly for the other broad classes. Note that the absence of an event is denoted by a bar. In addition, we have used the fact that the presence of the event sonorant implies the presence of the event speech, that is,

$$P(SC|\text{sonorant}, \mathbf{x}_t) = P(SC|\text{sonorant}, \text{speech}, \mathbf{x}_t)$$

TABLE I
TRAINING OF PHONETIC FEATURE SVMs

Branch in hierarchy	class +1	class -1
p_1	silence	speech
p_2	sonorant	non-sonorant
p_3	sonorant consonant	vowel
p_4	stop burst	frication noise

In EBS, we train one binary SVM for each bifurcation in the hierarchy. The probabilities p_i are posterior probabilities and we calculate them from the output of SVMs as described in Section III. Therefore, for the recognition of five broad classes, only four binary SVMs are needed. In Table I, we show the classes that are trained against each other for building the four SVMs. A clear advantage of this system is that each class to be recognized does not have to be trained against all of the other classes. For example, the samples of V do not have to be trained against the samples of the classes - SC, Fr, ST and silence. Instead, given the hierarchy, the samples of V are trained against the samples of SC for the SVM that calculates p_3 and the samples of V and SC are trained against the samples of ST and Fr for the SVM that calculates p_2 , and so on. For each binary classifier in Table I, a comparable amount of training data for the two classes is available. Moreover, since all the classifiers are binary, the method overcomes the need to find good multi-class SVMs. Although a non-probabilistic hierarchy can be used to limit the number of classifiers to four, such an approach will not allow probabilistic segmentation, therefore, the errors at the phonetic feature level will not be corrected by language and duration constraints.

A stop burst is characterized by a period of closure of about 30 ms and a transient burst. At a frame step size of 5ms, we use $s = 6$ and $e = 3$ for the stop burst classifier. For all of the other classifiers, a single frame of speech is used. That is, $s = 0$ and $e = 0$. With no duration and language constraints, and by making the independence assumption across frames, the class label at time t is hypothesized by

$$\hat{w}_t = \arg \max_w P(w|\mathbf{x}_t)$$

$$\text{with } w \in \{V, SC, ST, Fr, SIL\}$$

The segmentation of the test signal is then found by collapsing consecutive identical class labels. Note that this is not a strictly frame-based system because (1) a certain number of adjoining frames are used for stop burst detection and (2) acoustic-phonetic knowledge is used to normalize some of the APs across time, across frequency or both [15].

Table II shows the APs used by each SVM classifier. Unlike the HMM based approach, each classifier uses only the APs that are required for the corresponding phonetic feature.

V. EXPERIMENTS AND RESULTS

The training of the SVMs was performed with 5000 samples for each of the classes in Table I that were selected randomly from the TIMIT training files. Less than 1000 utterances were

TABLE II

APs USED IN BROAD CLASS SEGMENTATION. ZCR : ZERO CROSSING RATE, f_s : SAMPLING RATE, F3 : THIRD FORMANT AVERAGE. E[A,B] DENOTES ENERGY IN THE FREQUENCY BAND [AHZ,BHZ]

Branch in Hierarchy (Phonetic Feature)	APs
p_1 (Silence)	(1) E[0,F3-1000], (2) E[F3, $f_s/2$], (3) ratio of spectral peak in [0,400Hz] to the spectral peak in [400, $f_s/2$]
p_2 (Sonorant)	(1) Probability of voicing [16], (2) ZCR, (3) ratio of spectral peak in [0,400Hz] to the spectral peak in [400, $f_s/2$], (4) ZCR of high pass filtered speech, (5) Ratio of E[0,F3-1000] to E[F3-1000, $f_s/2$], (6) E[100,400]
p_3 (SC)	(1) E[640,2800] and (2) E[2000,3000] normalized by nearest syllabic peaks and dips
p_4 (Plosive)	(1) Energy onset, (2) Energy offset, (3) E[0,F3-1000], (4) E[F3-1000, $f_s/2$]

used for the training of the SVMs. The RBF kernels were used for the features silence, sonorant and plosive. A linear SVM was used for sonorant consonant detection.

HMM experiments [17] were carried out for comparison purposes using HTK [10]. A 39 parameter set consisting of 12 mel-frequency cepstral coefficients (MFCCs) and energy with their delta and acceleration coefficients was used in the HMM broad classifier. All of the manner class models were context-independent, 3-state (excluding entry and exit states), left-to-right HMMs with diagonal covariance matrices and 8-mixture observation densities for each state. A skip transition was allowed from the first state to the third state in each model. All the 'sx' and 'si' files (a total of 3696 utterances) were used for training the HMM broad classifier.

A manner class segmentation system may not separate out two consecutive phonemes having the same manner representation. Therefore, for the purpose of scoring of both the HMM system and EBS, the reference phoneme labels from the TIMIT database were mapped to manner class labels [6], and consecutive identical broad class labels were mapped into one. The resulting manner class labels were used as the reference labels for scoring EBS as well as the HMM broad classifier using the scoring package from NIST [18]. Except for the stop consonants, broad class labels that were ten milliseconds or shorter were deleted from the hypothesis segmentations because these are generally inserted at the boundary of two phonemes due to transient effects. The results are shown in Table III. EBS shows a better segmentation performance than the HMM based system with three times less parameters and four times less training data. Figure 2 shows the spectrogram of a TIMIT test sentence and the broad class labels generated by EBS as well as the HMM system along with the TIMIT phoneme labels. It can be easily seen from the examples shown in the figure that EBS does a finer analysis of the spectrum. In particular, the HMM broad classifier is not able to separate out a stop when it is followed by a fricative, while EBS can make this distinction.

TABLE III

RESULTS OF BROAD CLASSIFICATION

	HMM	EBS
Parameters	MFCCs	APs
Number of parameters	39	13
% Correct	69.6	79.8
% Accuracy	64.9	68.1

VI. DISCUSSION AND FUTURE WORK

We have presented a method for broad class segmentation, but the method can be extended to phoneme recognition and bigger recognition tasks in different ways by using the complete phonetic feature hierarchy in a probabilistic manner. An example of the representation of the phoneme /n/ with the phonetic feature hierarchy appears below

$$\begin{aligned}
 P(/n/|\mathbf{x}_t) &= P(\text{speech}, \text{sonorant}, \text{SC}, \text{nasal}, \text{alveolar}|\mathbf{x}_t) \\
 &= P(\text{speech}|\mathbf{x}_t)P(\text{sonorant}|\text{speech}, \mathbf{x}_t) \\
 &\quad P(\text{SC}|\text{sonorant}, \mathbf{x}_t)P(\text{nasal}|\text{SC}, \mathbf{x}_t) \\
 &\quad P(\text{alveolar}|\text{nasal}, \mathbf{x}_t)
 \end{aligned}$$

Although, the frame-level classification method we have proposed for broad class segmentation can be extended to take into account the complete phonetic feature hierarchy, the drawback of such a frame based approach is that coarticulation and dynamic variations within phonemes will not be appropriately modeled.

We are extending EBS for phoneme and word recognition using an event-based acoustic-phonetic approach. In this method, landmarks (like the locations of the burst of a stop, of the syllabic dip in a sonorant consonant and the syllabic peak in a vowel) are analyzed for place and voicing phonetic features. A very high recognition accuracy of the place and voicing features of stop consonants and fricatives has been obtained [19], [20], [21] using knowledge based measurements on segmented broad classes. The parameters listed in similar research and upcoming acoustic-phonetic research can be used with SVMs at appropriate landmarks to obtain the posterior probabilities of the different phonetic features in the hierarchy. Multiple hypothesis broad class segmentations will be generated using Viterbi or a beam search algorithm, and the broad segments will then be analyzed to find the posterior probabilities of the place and voicing features corresponding to the broad class segments. For word and sentence level recognition, broad class segmentation paths will be constrained by a pronunciation model such as a finite state automata (FSA).

Figure 3 shows a FSA for the broad class representation SIL-Fr-V-SC-V-SC-SIL of the digit 'zero', corresponding to the pronunciation /z I r ow/. To find the probability of the digit 'zero' given an utterance, the best path through the FSA in Figure 3 can be found using the binary SVM classifiers for

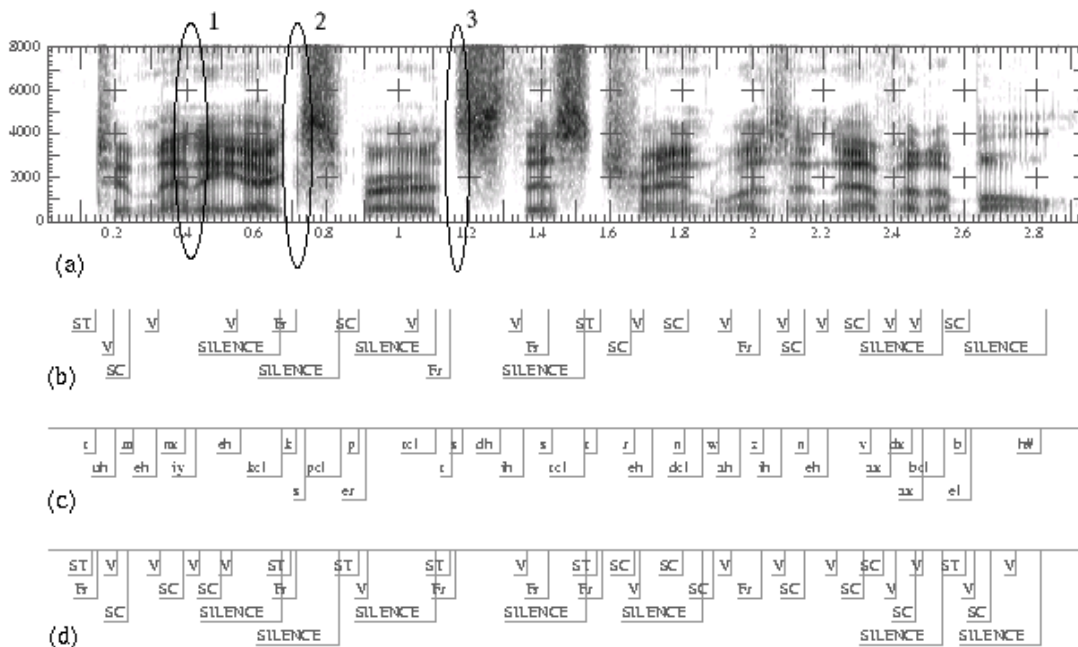


Fig. 2. (a) Spectrogram of TIMIT sentence "to many experts this trend was inevitable", (b) Labels generated by HMM method, (c) TIMIT phoneme labels, (d) Labels generated by EBS. In region 1, HMM broad classifier misses a short nasal segment. In regions 2 and 3, EBS could separate out a stop from a following fricative while the HMM recognizer misses the stop. In the EBS labels, the aspiration noise following the stop burst of /t/ is recognized as Fr. This is because the SVM model is built for the stop burst and not the complete stop region. Aspiration will be distinguished from frication as the hierarchical system is completed. It can also be seen that EBS gets the off-glide /y/ of the vowel /iy/ in the word 'many'

the broad classes. The probability of each of the phones /z/, /l/, /r/ and /ow/ will then be found by calculating the probabilities of the place, voicing and other features (for example, strident for fricatives) corresponding to each of the phones, as shown in Figure 3. For example, for /z/, the following probabilities will have to be estimated

- 1) $P(\text{voiced} | Fr, \mathbf{x}_t)$
- 2) $P(\text{strident} | \text{voiced}, Fr, \mathbf{x}_t)$
- 3) $P(\text{alveolar} | \text{strident}, \text{voiced}, Fr, \mathbf{x}_t)$

It is difficult to say whether the place features can be assumed to be independent of the voicing and manner features, or whether the voicing features can be assumed to be independent of the manner features. If highly discriminative APs that are correlates of the place features are found such that the discriminative ability of the APs is independent of the underlying manner and voicing features, we may be able to make such an assumption. For example, if APs are found that identify the feature alveolar across nasals, fricatives as well as stop consonants with sufficient accuracy, we may be able to assume,

$$\begin{aligned}
 P(\text{alveolar} | \text{strident}, \text{voiced}, Fr, \mathbf{x}_t) &= P(\text{alveolar} | \mathbf{x}_t) \\
 P(\text{alveolar} | \text{nasal}, \mathbf{x}_t) &= P(\text{alveolar} | \mathbf{x}_t) \\
 P(\text{alveolar} | \text{voiced}, \text{stop}, \mathbf{x}_t) &= P(\text{alveolar} | \mathbf{x}_t)
 \end{aligned}$$

Coarticulation is taken into account by the way the APs

for place and voicing are calculated. A typical example of an AP that takes context into account is the parameter Av-Ahi [22] suggested for distinguishing alveolar and labial stop consonants. Av-Ahi is calculated by subtracting the high frequency peak of the stop burst from the lowest frequency peak of the following vowel. Because coarticulation effects are implicitly modeled by the APs, there is no need in this system to build diphone and triphone models.

VII. CONCLUSION

We have shown that SVMs can be combined effectively with acoustic-phonetic knowledge both in terms of knowledge based APs and a phonetic feature hierarchy to provide a segmentation method for continuous speech. EBS does a very fine analysis of the speech spectrum and has a good capability of separating out transient sounds like stops. A probabilistic hierarchy allows the reduction of the number of support vector classifiers and avoids the need of using multi-class SVMs.

REFERENCES

- [1] V. Vapnik, "The Nature of Statistical Learning Theory", Springer Verlag, 1995.
- [2] P. Niyogi, "Distinctive Feature Detection Using Support Vector Machines", pp 425-428, ICASSP 1998.
- [3] J. Keshet, D. Chazan and B. Bobrovsky, "Plosive Spotting with Margin Classifiers", Eurospeech 2001.
- [4] P. Clarkson, P. J. Moreno, "On The Use Of Support Vector Machines For Phonetic Classification", ICASSP '99. [http:// cite-seer.nj.nec.com/clarkson99use.html](http://cite-seer.nj.nec.com/clarkson99use.html)

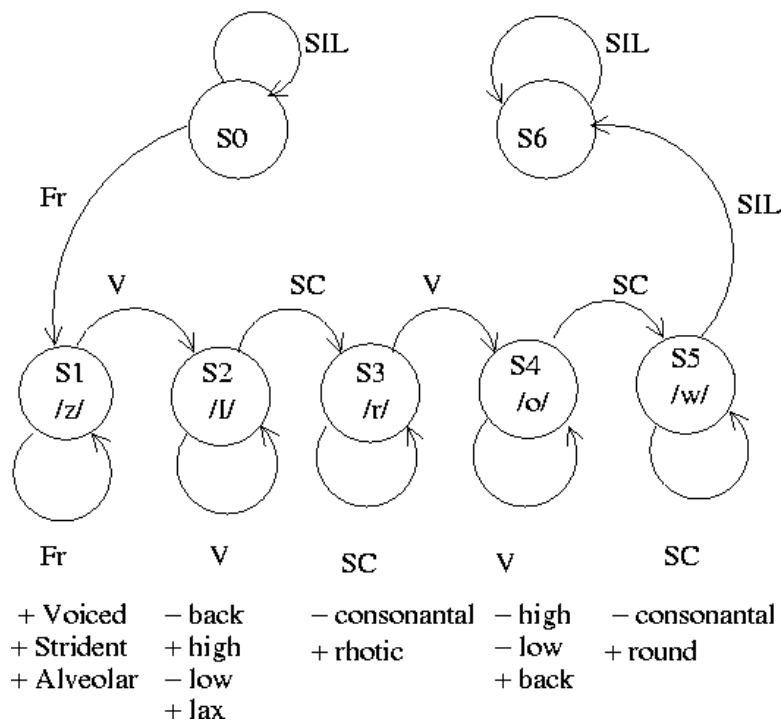


Fig. 3. A phonetic feature based language model for pronunciation of the digit 'zero'. FSA is shown for the broad class path SIL-Fr-V-SC-V-SC-SIL, with starting state S0 and end state S6. The place, voicing and other features (for example, strident) for which the posterior probabilities are required to be calculated for the phones /z/, /l/, /r/, /o/ and /w/ are shown beneath the states. For example, for /z/, the probabilities $P(\text{voiced}|Fr, \mathbf{x}_t)$, $P(\text{strident}|\text{voiced}, Fr, \mathbf{x}_t)$, $P(\text{alveolar}|\text{strident}, \text{voiced}, Fr, \mathbf{x}_t)$, need to be calculated to obtain the posterior probability of the word 'zero' given an utterance.

[5] H. Shimodaira, K. Noma, M. Nakai, S. Sagayama, "Support Vector Machine with Dynamic Time-Alignment Kernel for Speech Recognition", Eurospeech 2001

[6] A. Juneja and C. Espy-Wilson, "Segmentation of Continuous Speech Using Acoustic-Phonetic Parameters and Statistical Learning", in the proceedings of 9th International Conference on Neural Information Processing, Singapore, 2002, Volume 2, Page 726-730 .

[7] M. Halle and G. N. Clements, "Problem Book in Phonology", Cambridge, MA, MIT Press, 1983.

[8] J. B. Allen, "From Lord Rayleigh to Shannon: How do humans decode speech?", <http://auditorymodels.org/jba/PAPERS/ICASSP> .

[9] L. Rabiner, B. Juang, "Fundamentals of speech recognition", Prentice Hall, 1993.

[10] HTK documentation, <http://htk.eng.cam.ac.uk/>

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

[12] Ali, A. M. A., "Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition", Ph.D. Thesis, University of Pennsylvania, 1999.

[13] "TIMIT Acoustic -Phonetic Continuous Speech Corpus", National Institute of Standards and Technology Speech Disc 1 -1.1, NTIS Order No. PB91 -5050651996, October 1990

[14] T. Joachims, "Making large -Scale SVM Learning Practical", LS8-Report 24, Universita't Dortmund, LS VIII-Report, 1998.

[15] N. Bitar, "Acoustic Analysis and Modelling of Speech Based on Phonetic Features", PhD thesis, Boston University, 1997

[16] ESPS (Entropic Signal Processing System 5.3.1), Entropic Research Laboratory, <http://www.entropic.com>

[17] HMM experiments carried out at Speech Communication Lab by Om Deshmukh, <http://www.ece.umd.edu/omdesch/iconip2002.html>

[18] Speech Recognition Scoring Package (SCORE) Version 3.6.2, <http://www.nist.gov/speech/tools/>

[19] A. Juneja and C. Espy-Wilson, "An Event-Based Acoustic-Phonetic Approach for Speech Segmentation and E-Set Recognition", ICPhS 2003, Barcelona, Spain.

[20] A.M.A. Ali, J. V. Spiegel and P. Mueller, "An Acoustic-Phonetic Feature-based System for the Automatic Recognition of Fricative Consonants", ICASSP-98, vol. II, pp. 961-964, 1998.

[21] A. M. A. Ali, J. V. Spiegel and P. Mueller, "Automatic Detection and Classification of Stop Consonants using an Acoustic-Phonetic Feature-Based System", XIVth International Congress of Phonetic Sciences, pp. 1709-1712, 1999.

[22] K.S. Stevens, S. Manuel, and M. Matthies, "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production", Proceedings of the International Congress of Phonetic Sciences (1999).