SEGMENTATION OF CONTINUOUS SPEECH USING ACOUSTIC-PHONETIC PARAMETERS AND STATISTICAL LEARNING

Amit Juneja and Carol Espy-Wilson
ECE Department, University of Maryland, College Park, MD 20742, USA
http://www.ece.umd.edu/~juneja

ABSTRACT

In this paper, we present a methodology for combining acoustic-phonetic knowledge with statistical learning for automatic segmentation and classification of continuous speech. At present we focus on the recognition of broad classes - vowel, stop, fricative, sonorant consonant and silence. Judicious use is made of 13 knowledge-based acoustic parameters (APs) and support vector machines (SVMs). It has been shown earlier that SVMs perform comparable to hidden Markov models (HMMs) for detection of stop consonants. We achieve performance on segmentation of continuous speech better than the HMM based approach that uses 39 cepstrum-based speech parameters.

1. INTRODUCTION

There is strong evidence that human speech recognition (HSR) starts with a bottom-up analysis [1], and then later context is integrated into the recognition process. Present state-of-the-art automatic speech recognition (ASR) systems are top-down [2,3]. That is, the process starts by taking a dictionary of words and constituent phonemes. Each entry in the dictionary is a word with one or more sequences (pronunciations) of constituent phonemes. Hidden Markov models (HMMs) are built for each phone (monophone model) or triphone (triphone models). For the purpose of recognition, the best path through a lattice of words is found and the corresponding sequence of words is chosen as the most likely sequence. The front ends of ASR usually consist of mel-frequency cepstral coefficients (MFCCs) or perceptual linear predictive coefficients (PLPs).

We are developing an acoustic-phonetic approach to speech recognition in which speech is first segmented into broad classes (*vowel, stop, fricative, sonorant consonant* and *silence)*. These manner based segments are then analyzed for place of articulation to decide upon the constituent phonemes. Acoustic-phonetic approaches are bottom-up, but they have been overpowered by statistical pattern recognition approaches primarily because (1) acoustic-phonetic approaches have used hard coded decision rules that are not easy to adapt and (2) mapping of phonemes to sentences is a difficult task. On the other hand, since an acoustic-phonetic approach to recognition involves the explicit extraction of linguistic information that is combined for recognition, it is relatively straightforward to pinpoint the cause of recognition errors. This diagnosis is typically difficult in HMM-based systems where it is hard to determine if errors are due to failure of the pattern matcher or ill-represented speech information.

Our goal in this paper is to develop a system that combines the strengths of an acoustic-phonetic approach and statistical pattern matching. In particular, we have developed an adaptable and modular system where it is easy to assess the full system as well as the components for errors. Phonetic feature theory provides a hierarchical framework [6] and support vector machines (SVMs) provide the methodology for combining the speech knowledge. The success of SVMs has been demonstrated for the problem of detection of stop consonants [5]. We concentrate on the intensive use of knowledge-based parameters with SVMs for automatic segmentation of speech.

2. DATABASE

The TIMIT database [4] was used as a corpus of labeled speech data. Phonetically rich 'sx' and 'si' sentences from all the eight dialect regions in the training set were used for training and development, and the 'si' sentences from all the dialect regions in the test set (spoken by an independent set of speakers) were used for testing.

3. SUPPORT VECTOR MACHINES

SVM [15,16] is a statistical learning method for regression and pattern classification. While learning from data, SVM performs *structural risk minimization* (SRM) unlike the classical adaptation methods that minimize training error in a specific norm. For the two-class pattern classification problem, SVM finds a

decision hypersurface $d(x)$, where the vector x belongs to the space of samples, of the following form

$$d(x) = \sum_{i=1}^{l} y_i a_i K(x, x_i)$$

The support vectors (SVs) $\{x_i\}_{i=1}^{l}$, and the weights a_i are found by using quadratic optimization methods and the training data. y_i are the class labels of the support vectors that take the value +1 or –1 depending upon the class. The kernel $K(x, x_i)$ is a function of the dot product of the vectors x and x_i. The kernel depends on the type of the hypersurface $d(x)$. The kernels used in this project are shown in Table 1 along with the type of the hypersurface. For a test vector x, the class is determined by the sign of $d(x)$. The experiments in this project were carried using the SVM Light toolkit [8], which provides very fast training of SVMs.

Hypersurface type	Kernel	Kernel specific parameter
Linear	$(x.x_i + 1)$	None
Polynomial of degree d	$(x.x_i + 1)^d$	d
Gaussian Radial base function (RBF)	$\exp(-\gamma \mid x - x_i \mid^2)$	γ

Table 1 : SVM kernels and their corresponding kernel-specific parameters

4. METHOD

Our event-based speech recognition system (EBS) has four modules – an acoustic-phonetic knowledge based parameter extraction front-end, a statistical learning module, multi-class decision module and a language modeling module. In this paper, we concentrate on the first three modules for the task of segmentation of speech into five broad classes (Figure 1). The front end generates 13 APs that are acoustic correlates of the manner phonetic features [10] – *syllabic*, *sonorant*, *noncontinuant*, *obstruent* and silence. Using these acoustic correlates, speech is segmented into the broad classes mentioned before. The different phonemes of English that lie in each of the manner classes are shown in Table 2. This classification of phonemes is not strict since the surface realization can be significantly different from its canonical form due to coarticulatory effects and weakening processes (lenition).

Manner	Phoneme
Vowel	ih, eh, ae, aa, ah, ao, uh, ah, ax, ih, ax, axr-h, en, em, eng, el, er
Vowel followed by sonorant consonant	iy, ey, ay, ow, oy, aw
Sonorant consonant	w, l, r, y, m, n, ng, nx, dx, hv
Fricative	s, sh, f, th, hh, z, zh, v, dh
Stop followed by fricative	jh, ch
Stop	b, d, g , p, t, k

Table 2: Broad manner classification of English phonemes

Speech is analyzed every 5ms with a 10ms Hanning window (5ms overlap). Knowledge-based parameters are extracted from both the time waveform and the spectrum of the signal. A classifier is built for each of the classes – vowel, sonorant consonant, fricative, stop and silence. In practice, the classifier for sonorancy is used in place of classifier of vowel because vowels and sonorant consonants are both sonorants and they are distinguished by the classifier for sonorant consonants. Each classifier operates on a frame of speech and takes the acoustic parameters for that frame and in some cases, a particular number of adjoining frames. Not all acoustic parameters are used by each classifier. The parameters for each classifier are chosen on the basis of knowledge. The output of each classifier is mapped to a probability measure, that is, the a posteriori probability of the manner class. A very crude form of this mapping is used in the current set of experiments. The SVM outputs are clipped in the range [-1,1], then scaled and translated to the range [0,1].

At each frame, the probability outputs of the classifiers are compared and the maximum is chosen. A speech segment is then hypothesized by a region in which the output of a single classifier remains the maximum. Note that we have only outlined the system for broad class segmentation. For phoneme and sentence recognition, there will be a classifier for each of the 20 phonetic features that are known to be sufficient to describe the sounds in all the languages in the world [6]. We now discuss the design and the parameter selection of the classifiers.

4.1 APs

Table 3 shows the APs used for the detection of each manner class. Except for stop detection, parameters only from the current analysis frame are fed to the SVM. Stops are characterized by a period of silence (closure) followed by a sudden release in energy (onset) and then

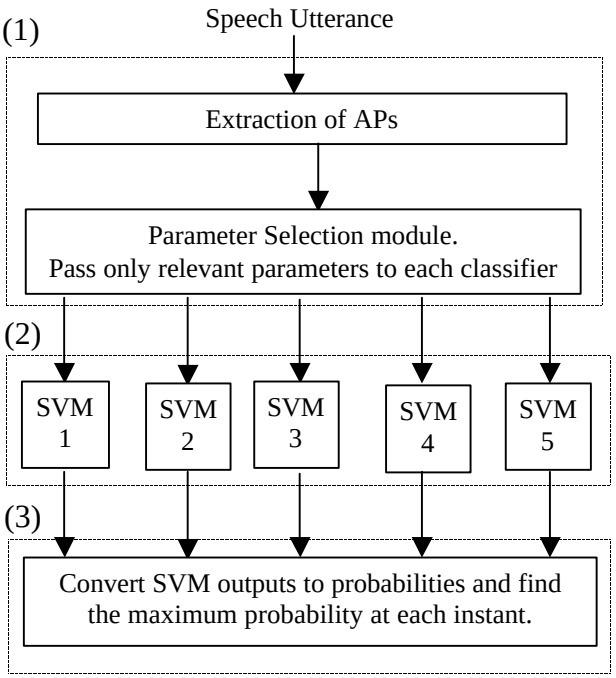

Figure 1 : The three modules of EBS – (1) front end, (2) pattern recognition and (3) multi-class decision. SVM 1 : vowel detection, SVM 2 : sonorant consonant detection, SVM 3 : fricative detection, SVM 4 : stop detection, SVM 5 : silence detection

a sudden fall in energy (offset). Therefore, for detection of stops information not only from one frame but adjoining frames is required. Especially, there is about 30ms of silence before stop bursts [11], so we use 6 frames preceding the analysis frame and 3 frames following the analysis frame for the detection of stop burst.

4.2 SVM kernel selection

We trained three different SVMs – linear, polynomial and RBF – for the detection of each manner class. Sonorant frames were trained against all non-sonorant frames including frication, silence, and stops. 30,000 frames of speech were selected for each class randomly from the TIMIT training data, from both male and female utterances. The Xi-Alpha estimates [8,9] of the error bound provided by the learning process and the number of support vectors for each machine is shown in Table 4.

We choose RBF kernel with $\gamma = 0.01$ for speech segmentation experiments because of lowest error bound estimate of 10.86%. Similar analysis was carried out for other manner classes. The choice of SVM kernel and error bound estimate for each class is shown in Table 5.

Manner Class	Parameters
Sonorant	(1) Probability of voicing [7], (2) ZCR, (3) ratio of spectral peak in [0,400] to the spectral peak in [400, SF/2], (4) ZCR of high pass filtered speech, (5) Ratio of E[0, F3-1000] to E[F3-1000, SF/2], (6) E[100,400]
Stop	(1) Energy onset (2) Energy offset (3) E[0,F3] (4) E[F3, SF/2]
Fricative	Same as sonorant parameters, and E[F3, SF/2]
Sonorant consonant	(1) E[640, 2800], (2) E[2000,3000]
Silence	(1) E [0,F3], (2) E[F3,SF/2], (3) ratio of spectral peak in [0,400] to the spectral peak in [400, SF/2]

Table 3 : APs used for detection for each manner class. ZCR : zero crossing rate, SF : sampling frequency, F3 : third formant of the speaker, E[a, b] denotes energy in the frequency band [aHz, bHz].

Kernel	Kernel-specific Parameter	Number of SVs	Error estimate (%)
Linear	-	9874	16.35
Polynomial	d = 2	10133	16.66
Polynomial	d = 3	9727	16.10
RBF	$\gamma = 0.05$	27474	16.79
RBF	$\gamma = 0.01$	13902	10.86
RBF	$\gamma = 0.005$	10458	11.47

Table 4 : Training record of sonorancy SVM. Not all values of γ are shown.

The error bound estimate in detection of sonorant consonants is high because boundaries between vowels and sonorant consonants are not well defined and there is a lot of overlap in the training data. This does not harm so much because even if only the central regions of the sonorants consonants are detected that would suffice for the purpose of phoneme recognition. This may, though, cause insertions of sonorant consonants in the vowel regions with weak energies but that problem can be solved by using temporal parameters for sonorant consonants [14].

4.3 Analysis of SVM outputs

Figure 2 shows two of the parameters for sonorancy detection – ZCR and ratio of E[0, F3-1000] to E[F3-1000, SF/2] – plotted against a speech spectrum. Also shown is the output of sonorancy SVM converted to probability estimate. Sonorant regions have low ZCR and large designated ratio of energies. High values of

Manner Class	Kernel	Kernel-Parameter	Error-bound estimate (%)
Sonorant	RBF	ã =0.01	10.86
Stop	RBF	ã=0.001	7.41
Fricative	RBF	ã = 0.008	14.31
Sonorant consonant	Linear	none	49.70
Silence	RBF	ã = 0.001	14.92

Table 5 : SVM kernel selection for different manner classes.

probability are obtained in the sonorant regions and low values are obtained in the non-sonorant regions as per expectations. Figure 2 illustrates the ease in which fault can be found with the system. The oval region in the spectrum is a /t/ and is not a sonorant region but as shown by the arrow, we get a high probability of sonorancy in that region.

This error can be easily explained by presence of low ZCR (compared to fricatives) and high E[0,3000Hz] which is characteristic of sonorants. That is, the problem lies in the parameters. It can be fixed by checking if the high energy in the low frequency band is periodic or aperiodic [14], that is, by modulating the low frequency energy by the periodicity in the low frequency bands. If the speech is degraded, similar plots can be obtained to see if it is the parameters that are not behaving in line with their physical significance. However, if in degraded speech, the parameters are behaving well but the recognition is not good, outputs of different SVMs can be plotted with the spectrogram to find which SVMs are going wrong.

4.4 HMM experiments

HMM experiments [17] were carried out using HTK [3]. 39-parameter set consisting of 12 MFCCs and energy with their delta and acceleration coefficients were used in the HMM broad classifier. All 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.

5. RESULTS AND DISCUSSION

A manner class segmentation system may not separate out two consecutive phonemes having the same manner representation. Therefore, for the purpose of scoring, the reference phoneme labels from the TIMIT database were mapped to manner class labels with the mappings listed in Table 1, and the consecutive identical manner labels were collapsed into one. The resulting manner class labels were used as the reference labels for scoring EBS as well as the HMM broad classifier. EBS with 13 APs showed performance better than HMMs with 39 cepstrum-based parameters. The results are shown in Table 6.

	HMM	EBS
Parameters	MFCCs	APs
Number of parameters	39	13
% Correct	69.6	82.4
% Accuracy	64.9	68.3

Table 6 : Results of broad classification

The wide gap in the correctness and accuracy of EBS is because of a higher number of insertions, primarily, of sonorant consonants and stops. Stop insertions normally occur at the onset of vowels and strong fricatives following a period of silence. Sonorant consonant insertions occur at the weak beginning and end of vowels. These insertions may be corrected by using temporal parameters [14] as well as designing more discriminative parameters.

6. CONCLUSION AND FUTURE WORK

We have seen that statistical learning can be applied successfully with the knowledge of acoustic-phonetics for segmentation of speech with performance comparable to HMM systems. The recognition method makes it easy to find the source of error in the system. The system can be easily retrained for any new set of parameters or for recognition of other languages. The work will be extended to complete phoneme recognition in the future. Neural networks that perform equally well may replace SVMs where the number of support vectors is too high for real-time operation of EBS. Better methods of conversion of SVM outputs to probabilities [13] will be applied. These parameters will be used and tested with EBS in noise robust conditions. At present the learning in EBS is supervised, that is, the system requires time-aligned labeled data for training. In the future we will explore the possibility of unsupervised learning for the system.

7. REFERENCES

[1] J. B. Allen, “From Lord Rayleigh to Shannon: How do humans decode speech?”, http://auditorymodels.org/jba/PAPERS/ICASSP/

[2] L. Rabiner, B. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, 1993

[3] HTK documentation, http://htk.eng.cam.ac.uk/

[4] ``TIMIT Acoustic-Phonetic Continuous Speech Corpus", National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-5050651996, October 1990
[5] P. Niyogi, "Distinctive Feature Detection Using Support Vector Machines", pp 425-428, ICASSP 1998.
[6] M. Halle and G. N. Clements, "Problem Book in Phonology", Cambridge, MA, MIT Press, 1983.
[7] ESPS (Entropic Signal Processing System 5.3.1), Entropic Research Laboratory, http://www.entropic.com
[8] T. Joachims, "Making large-Scale SVM Learning Practical", LS8-Report 24, Universität Dortmund, LS VIII-Report, 1998.
[9] T. Joachims, "Estimating the Generalization Performance of a SVM Efficiently", Proceedings of the International Conference on Machine Learning, Morgan Kaufman, 2000.
[10] N. Bitar, "Acoustic Analysis and Modelling of Speech Based on Phonetic Features", PhD thesis, Boston University, 1997
[11] K. Stevens, "Acoustic-Phonetics", MIT Press, ISBN: 026219404X, 1999
[12] T. Briscoe, "Lexical access in connected speech recognition", P98-1011, Computational Linguistics, Association for Computational Linguistics, 1989.
[13] .T. Kwok, "Moderating the outputs of support vector machine classiers". IEEE Transactions on Neural Networks, 10:1018-1031, 1999.
[14] A. Saloman, and C. Espy-Wilson, "Automatic Detection of Manner Events for a Knowledge-Based Speech Signal Representation", Proc.of Eurospeech, Sept. 1999, Budapest Hungary, pp. 2797-2800.
[15] V. Vapnik, "The Nature of Statistical Learning Theory", SpringerVerlag, 1995.
[16] V. Kecman, "Learning and Soft Computing – Support Vector Machines, Neural Networks and Fuzzy Logic models", The MIT Press, Cambridge, MA, 2001
[17] HMM experiments carried out at Speech Communication Lab by Om Deshmukh, http://www.ece.umd.edu/iconip2002.html

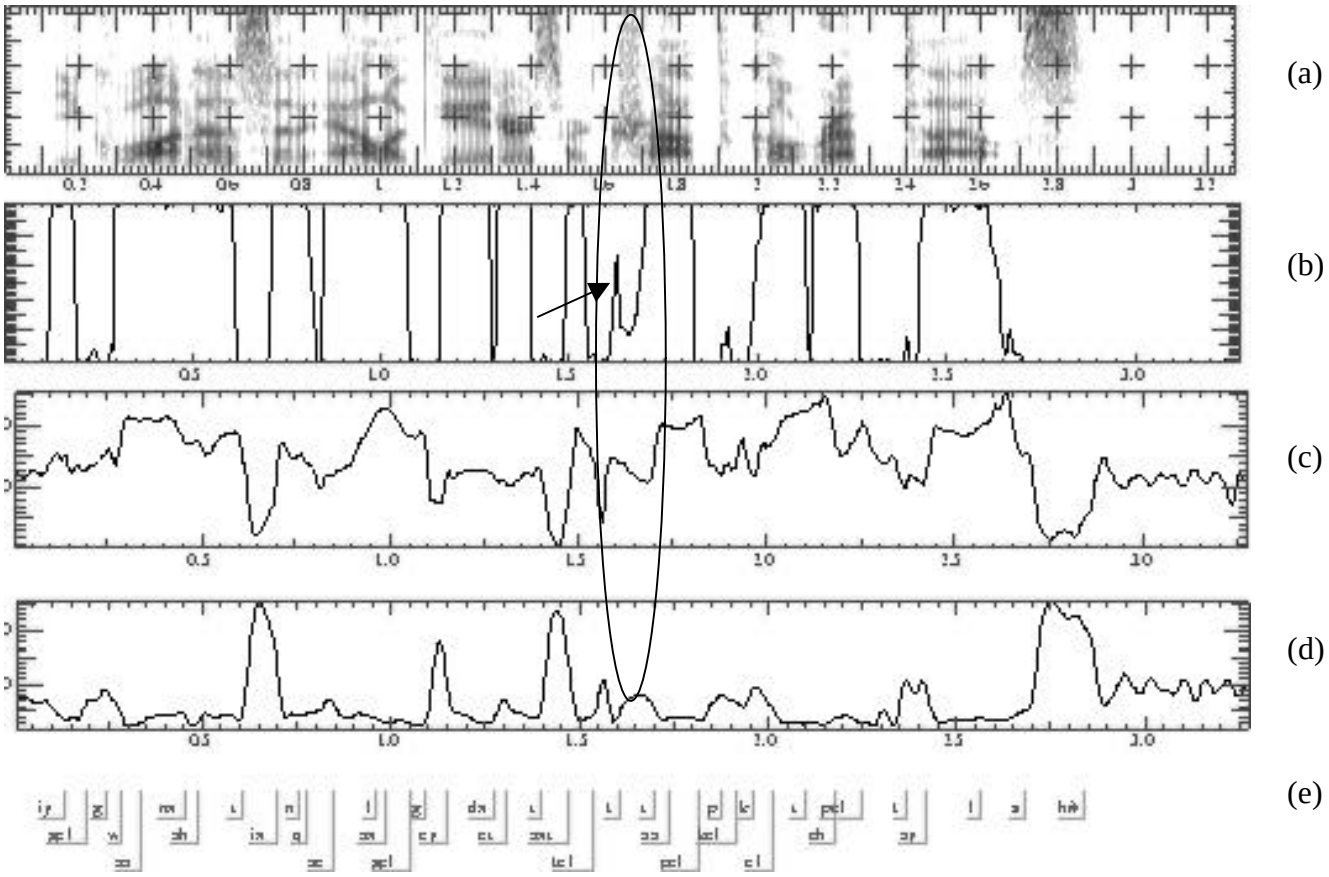

Figure 2 : Spectrogram of the utterance "Iguanas and alligators are tropical reptiles". (a) Spectrogram, (b) SVM a posteriori probability of sonorancy, (c) Ratio of E[0,F3-1000] to E[F3-1000,SF/2], (d) Zero crossing rate (e) Phoneme labels from TIMIT database. The phoneme in the oval region is /t/