# A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition[a)]

Amit Juneja[b)] and Carol Espy-Wilson
*Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742*

(Received 15 November 2007; accepted 20 November 2007)

A probabilistic framework for a landmark-based approach to speech recognition is presented for obtaining multiple landmark sequences in continuous speech. The landmark detection module uses as input acoustic parameters (APs) that capture the acoustic correlates of some of the manner-based phonetic features. The landmarks include stop bursts, vowel onsets, syllabic peaks and dips, fricative onsets and offsets, and sonorant consonant onsets and offsets. Binary classifiers of the manner phonetic features—syllabic, sonorant and continuant—are used for probabilistic detection of these landmarks. The probabilistic framework exploits two properties of the acoustic cues of phonetic features—(1) sufficiency of acoustic cues of a phonetic feature for a probabilistic decision on that feature and (2) invariance of the acoustic cues of a phonetic feature with respect to other phonetic features. Probabilistic landmark sequences are constrained using manner class pronunciation models for isolated word recognition with known vocabulary. The performance of the system is compared with (1) the same probabilistic system but with mel-frequency cepstral coefficients (MFCCs), (2) a hidden Markov model (HMM) based system using APs and (3) a HMM based system using MFCCs. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2823754]

PACS number(s): 43.72.Ne, 43.72.Bs, 43.72.Ar [ADP] Pages: 1154–1168

## I. INTRODUCTION

In a landmark-based automatic speech recognition (ASR) system, the front-end processing (or low-level signal analysis) involves the explicit extraction of speech-specific information. This speech-specific information consists of the acoustic correlates of the linguistic features (Chomsky and Halle, 1968) which comprise a phonological description of the speech sounds. The processing occurs in two steps. The first step consists of the automatic detection of acoustic events (also called landmarks) that signal significant articulatory changes such as the transition from a more open to a more closed vocal tract configuration and vice versa (i.e., changes in the manner of articulation), a sudden release of air pressure and changes in the state of the larynx. There is evidence that the auditory system responds in a distinctive way to such acoustic events (e.g., Delgutte and Kiang, 1984). The second step involves the use of these landmarks to extract other relevant acoustic information regarding place of articulation that helps in the classification of the sounds spoken. Given the extensive variability in the speech signal, a complete ASR system would integrate this front-end processing with a lexical access system that handles pronunciation variability and takes into account prosody, grammar, syntax and other higher-level information.

State-of-the-art ASR systems are based on hidden Markov modeling (HMM) and the standard parametrization of the speech signal consists of Mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives (Rabiner and Juang, 1993; Young *et al.*, 2006). The HMM framework assumes independence of the speech frames so that each one is analyzed and all of the MFCCs are looked at in every frame. In contrast, a landmark-based approach to speech recognition can target level of effort where it is needed. This efficiency can be seen in several ways. First, while each speech frame may be analyzed for manner-of-articulation cues resulting in the landmarks, analysis thereafter is carried out only at significant locations designated by the landmarks. This process in effect takes into account the strong correlation among the speech frames. Second, analysis at different landmarks can be done with different resolutions. For example, the transient burst of a stop consonant may be only 5 ms long. Thus, a short temporal window is needed for analysis. On the other hand, vowels which are considerably longer (50 ms for a /schwa/ to 300 ms for an /ae/) need a longer analysis window. Third, the acoustic parameters (APs) used to extract relevant information will depend upon the type of landmark. For example, at a burst landmark, appropriate APs will be those that characterize the spectral shape of the burst (maybe relative to the vowel to take into account contextual influences) to distinguish between labial, alveolar and velar stops. However, at a vowel landmark, appropriate APs will be those that look at the relative spacing of the first three formants to determine where

[a)] Portions of this work have appeared in "Segmentation of Continuous Speech Using Acoustic-Phonetic Parameters and Statistical Learning," International Conference on Neural Information Processing, Singapore, 2002, and "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," International Joint Conference on Neural Networks, Portland, Oregon, 2003, and "Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition," in the proceedings of From Sound to Sense: 50+ Years of Discoveries in Speech Communication, June 11–13, 2004, MIT, Cambridge, MA.

[b)] Author to whom correspondence should be addressed. Electronic mail: amjuneja@gmail.com

 

the vowel fits in terms of the phonetic features *front*, *back*, *high* and *low*.

Another prominent feature of a landmark ASR system is that it is a tool for uncovering and subsequently understanding variability. Given the physical significance of the APs and a recognition framework that uses only the relevant APs, error analysis often points to variability that has not been accounted for. For example, an early implementation of the landmark-based ASR system (Bitar, 1997) used zero crossing rate as an important measure to capture the turbulent noise of strident fricatives. The zero crossing rate will not be large, however, during a voiced strident fricative /z/ when it contains strong periodicity. In this case, the high-frequency random fluctuations are modulated by the low-frequency periodicity. This situation occurs when the /z/ is produced with a weakened constriction so that the glottal source is comparatively stronger (and, therefore, the supraglottal source is weaker) than it is during a more canonically produced /z/. Spectrographically, a weakened /z/ shows periodic formant structure at low frequencies like a sonorant consonant and some degree of turbulence at high frequencies like an obstruent. This understanding led to the development of an aperiodicity/periodicity/pitch (APP) detector which, along with fundamental frequency information, provides a spectro-temporal profile of aperiodicity and periodicity (Deshmukh *et al.*, 2005). The APP detector, however, was not used in this work due its computational requirements.

A significant amount of work has gone into understanding the acoustic correlates of the linguistic features (Stevens, 2002). Studies have shown that the acoustic correlates of the phonetic features can be reliably and automatically extracted from the speech signal (Espy-Wilson, 1987; Bitar, 1997; Ali, 1999; Carbonell *et al.*, 1987; Glass, 1984; Hasegawa-Johnson, 1996) and that landmarks can be automatically detected (Bitar, 1997; Salomon, 2000; Ali, 1999; Liu, 1996). Stevens (2002) has laid out a model for lexical access based on acoustic landmarks and phonetic features. However, to date, no one has implemented a complete ASR system based on a landmark approach. The previous landmark-detection systems (Bitar, 1997; Salomon, 2000; Ali, 1999; Liu, 1996) performed well, but they lacked a probabilistic framework for handling pronunciation variability that would make the systems scalable to large-vocabulary recognition tasks where higher-level information has to be integrated. For example, it could not be demonstrated in these systems that a voiced obstruent realized as a sonorant consonant will ultimately be recognized correctly due to higher-level constraints. These systems were primarily rule based and it has been pointed out (Rabiner and Juang, 1993) that in rule-based systems, the difficulty in the proper decoding of phonetic units into words and sentences increases sharply with an increase in the rate of phoneme insertion, deletion and substitution. In this work, a probabilistic framework is developed that selectively uses knowledge-based APs for each decision and it can be constrained by a high-level pronunciation model of words and probability densities of durations of phonetic units. Since recognition can be constrained by higher-level knowledge, the system does not have to decode phonetic units into words in a separate step.

Probabilistic frameworks exist for segment-based (Glass *et al.*, 1996; Zue *et al.*, 1989; Halberstadt, 1998) and syllable-based (Chang, 2002) ASR. But these systems are not targeted at selectively using knowledge-based acoustic correlates of phonetic features for detection of landmarks or for place of articulation detection. Many HMM-based approaches to speech recognition have used knowledge-based APs (Bitar and Espy-Wilson, 1996; Deshmukh *et al.*, 2002; Hosom, 2000), or the concept of phonetic features (Deng and Sun, 1994; Kirchhoff, 1999; Eide *et al.*, 1993). However, these were not landmark-based methods in that they did not involve an initial step of segmenting or detecting events in speech.

In this paper, a probabilistic framework for a landmark-based ASR system called event-based system (EBS) (Bitar, 1997) is presented. The focus of this paper is on the implementation and performance of the probabilistic landmark detection module of the framework. The framework was discussed in brief earlier (Juneja and Espy-Wilson, 2004) but it was not sufficiently detailed because of the limited space available. Initial results during the development of the probabilistic landmark detection system were reported earlier (Juneja and Espy-Wilson, 2002, 2003) but these only involved statistical classifiers for phonetic features and did not involve the probabilistic scoring with duration constraints. Because of the lack of probabilistic duration constraints these initial systems resulted in numerous insertions of segments of very small durations (5–10 ms) and such small segments of continuant sounds had to be removed to get a good recognition accuracy. Also because of the lack of probabilistic scoring, these systems could not be constrained by pronunciation models. In this work, the complete probabilistic framework for landmark detection is described in detail along with the details of the experiments.

The performance of the framework is demonstrated with the broad-class recognition and landmark detection task using the TIMIT database (NIST, 1990) and a vocabulary-constrained broad class recognition task on isolated digits using the TIDIGITS database (LDC, 1982). These recognition tasks require only the first step of the front-end processing where information is sought regarding only the manner features. For comparison, a traditional HMM-based recognition system is implemented. In one case, the traditional MFCCs and their first and second derivatives serve as input to the HMM system. In another implementation, the same APs used in EBS serve as input to the HMM system. Finally, the APs in the EBS system are replaced with the MFCCs. This four-way comparison allows the evaluation of the effectiveness of not only the probabilistic landmark framework as compared to the statistical HMM system, but also the knowledge-based APs with the MFCCs.

## II. METHOD

### A. Overview

Figure 1 shows a block diagram of EBS and it highlights the part of EBS that is the focus of this paper. To start the landmark detection process, the knowledge-based APs (Bitar, 1997) shown in Table I for each of the phonetic features—

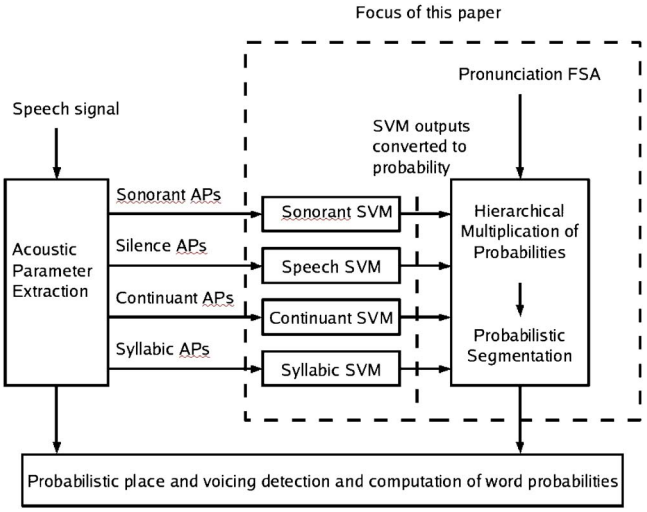


FIG. 1. (Color online) Overview of the landmark-based speech recognition system.

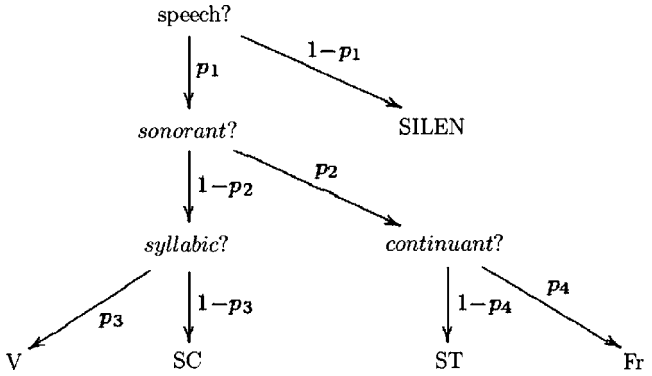


FIG. 2. Probabilistic Phonetic Feature Hierarchy.

*sonorant*, *syllabic*, *continuant*—and silence are automatically extracted from each frame of the speech signal. Then, a support vector machine (SVM) (Vapnik, 1995) based binary classifier is applied at each node of the hierarchy shown in Fig. 2 such that only the relevant APs for the feature at that node serve as input to the classifier. Probabilistic decisions obtained from the outputs of SVMs are combined with class dependent duration probability densities to obtain one or more segmentations of the speech signal into the broad classes—vowel (V), fricative (Fr), sonorant consonant (SC—including nasals semivowels), stop burst (ST) and silence (SILEN—including stop closures). A segmentation is then used along with the knowledge-based measurements to deterministically find landmarks related to each of the broad class segments. For a fixed vocabulary, segmentation paths can be constrained using broad class pronunciation models.

The phonetic feature hierarchy shown in Fig. 2 is the upper part of a complete hierarchy that has manner features at the top, place features at the lower nodes and phonemes at the lowest level. Several studies have provided evidence for a hierarchical organization of the phonetic features (e.g., Clements, 1985). Probabilistic hierarchies with phonemes at the terminal nodes have been used before in speech recognition (Halberstadt, 1998; Chun, 1996) where the use of such hierarchies occurs after an initial segmentation step. EBS uses the hierarchy as a uniform framework for obtaining manner-based landmarks and place and voicing feature detection. The complete hierarchy used by EBS is shown in Juneja, 2004.

Figure 3 shows the two kinds of landmarks that EBS is expected to extract (the landmark locations in this figure are hand marked)—abrupt landmarks and nonabrupt landmarks. In the previous implementation of EBS (Bitar, 1997), the maxima and minima of the APs $E[640\ \text{Hz}, 2800\ \text{Hz}]$ and $E[2000\ \text{Hz}, 3000\ \text{Hz}]$ were used in a rule-based system to obtain the nonabrupt landmarks that occur at syllabic peaks and syllabic dips. Thresholds on energy onsets and energy offsets were used to obtain the abrupt stop burst landmarks. Figure 3 shows that $E[640\ \text{Hz}, 2800\ \text{Hz}]$ has maxima in the vowel nuclei at the syllabic peak landmarks and minima in the sonorant consonant regions at the syllabic dip landmarks. Spurious peaks or dips caused insertions and peaks or dips that are not fully realized caused deletions in this system. There was no way to recover from such errors.

The presented framework can be viewed as a probabilistic version of the system in (Bitar, 1997) as it finds the landmarks using the following two steps:

1. The system derives multiple probabilistic segmentations from statistical classifiers (that use relevant APs as input) taking into account the probability distributions of the durations of the broad class segments. The probabilistic duration models penalize the insertions of very small broad class segments (for example, 5–10-ms-long vowels) by assigning low duration probabilities to such sounds.

TABLE I. APs used in broad class segmentation. $f_s$: sampling rate F3: third formant average [a,b]: frequency band [aHz,bHz], E[a,b]: energy in the frequency band [aHz,bHz].

| Phonetic Feature | APs |
|---|---|
| Silence | (1) $E[0, F3\text{-}1000]$, (2) $E[F3, f_s/2]$, (3) ratio of spectral peak in [0,400 Hz] to the spectral peak in $[400, f_s/2]$, (4) Energy onset (Bitar, 1997) (5) Energy offset (Bitar, 1997) |
| *sonorant* | (1) $E[0, F3\text{-}1000]$, (2) $E[F3, f_s/2]$, (3) Ratio of $E[0, F3\text{-}1000]$ to $E[F3\text{-}1000, f_s/2]$, (4) $E[100,400]$ |
| *syllabic* | (1) $E[640, 2800]$ (2) $E[2000, 3000]$ (3) Energy peak in [0,900 Hz] (4) Location in Hz of peak in [0,900 Hz] |
| *continuant* | (1) Energy onset (Bitar, 1997), (2) Energy offset (Bitar, 1997), (3) $E[0, F3\text{-}1000]$, (4) $E[F3\text{-}1000, f_s/2]$ |

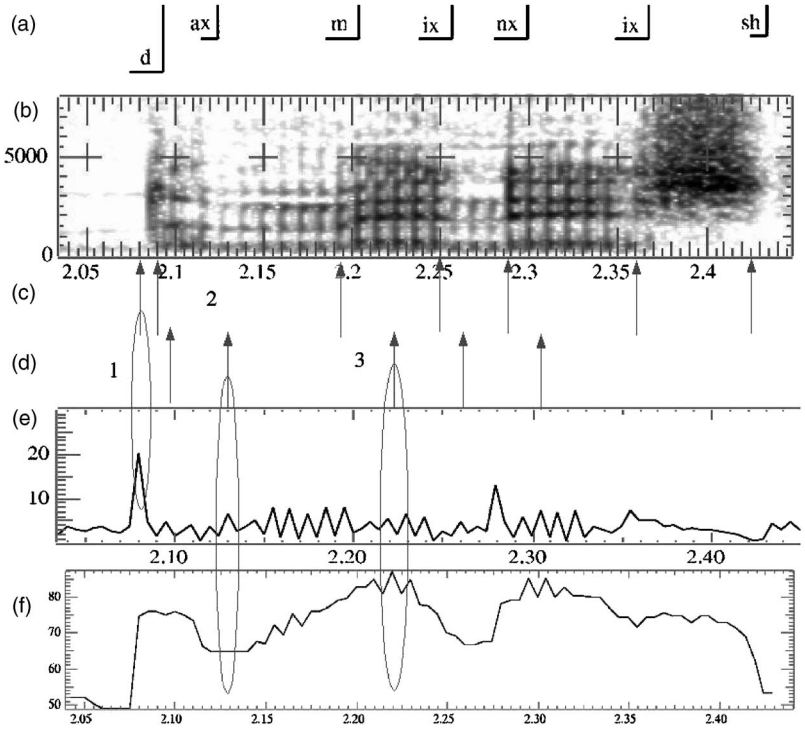


FIG. 3. Illustration of manner landmarks for the utterance "diminish" from the TIMIT database. (a) Phoneme Labels, (b) Spectrogram, (c) Landmarks characterized by sudden change, (d) Landmarks in stable regions of speech, (e) Onset waveform (an acoustic correlate of phonetic feature *−continuant*), (f) $E[640,2800]$ (an acoustic correlate of *syllabic* feature). The ellipses show how landmarks were obtained in (Bitar, 1997) using certain APs. Ellipse 1 shows the location of stop burst landmark for the consonant /d/ using the onset. Ellipse 2 shows the location of syllabic dip for the nasal /m/ using the minimum of $E[640,2800]$. Ellipse 3 shows that the maximum of the $E[640,2800]$ can be used to locate a syllabic peak landmark of the vowel /ix/.

2. The landmarks are then derived using the boundaries of the broad classes as abrupt landmarks and the maxima and minima of the AP $E[640\ \text{Hz},2800\ \text{Hz}]$ inside individual sonorant segments to get the nonabrupt landmarks. The global maximum inside the vowel segment is used to get the syllabic peak landmark and the global minima inside the sonorant consonant is used to get the syllabic dip landmark. Therefore, presence of multiple peaks or dips does not cause insertions at this step.

SVMs are used for the purpose of binary classification (although other classifiers, for example, neural networks or Gaussian mixture models could be used as well) of phonetic features because of their ability to generalize well to new test data after learning from a relatively small amount of training data. Additionally, SVMs have been shown to perform better than HMMs for phonetic feature detection in speech (Niyogi, 1998; Keshet *et al.*, 2001) and for phonetic classification from hand-transcribed segments (Clarkson and Moreno, 1999; Shimodaira *et al.*, 2001). The success of SVMs can be attributed to their property of large margin classification. Figure 4 shows two types of classifiers for linearly separable data: (1) a linear classifier without maximum margin and (2) a linear classifier with maximum margin. It can be seen from Fig. 4 that the maximum margin classifier is more robust to noise because a larger amount of noise (at least half of the margin for the samples shown) is required to let a sample point cross the decision boundary. It has been argued (Vapnik, 1995) that a maximization of the margin leads to the minimization of a bound on the test error. Mathematically, SVMs select a set of $N_{\text{SV}}$ support vectors $\{\mathbf{x}_i^{\text{SV}}\}_{i=1}^{N_{\text{SV}}}$ that is a subset of $l$ vectors in the training set $\{\mathbf{x}_i\}_{i=1}^{l}$ with class labels $\{y_i\}_{i=1}^{l}$, and find an optimal separating hyperplane $f(\mathbf{x})$ (in the sense of maximization of margin) in a high dimensional space $\mathcal{H}$,

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

The space $\mathcal{H}$ is defined by a linear or nonlinear kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ that satisfies the Mercer conditions (Burges, 1998). The weights $\alpha_i$, the set of support vectors $\{\mathbf{x}_i^{\text{SV}}\}_{i=1}^{N_{\text{SV}}}$ the

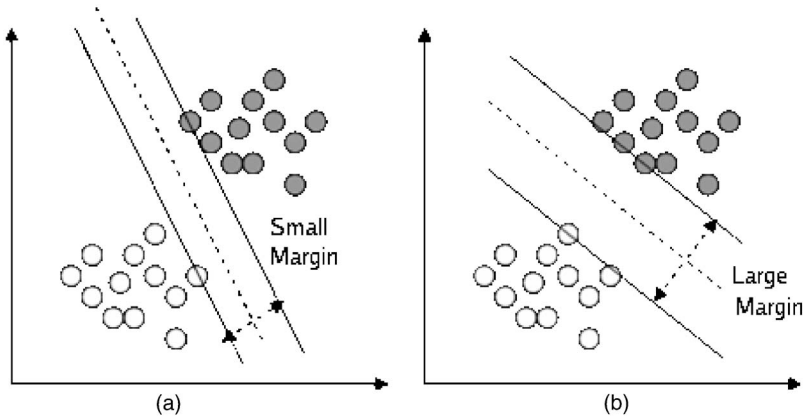


FIG. 4. (a) small margin classifiers, (b) maximum margin classifiers.

TABLE II. An illustrative example of the symbols $B$, $L$ and $U$.

| | /z/ | /ɪ/ | /r/ | /o/ | /w/ |
|---|---|---|---|---|---|
| $U\Rightarrow$ | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ |
| | *−sonorant* | *+sonorant* | *+sonorant* | *+sonorant* | *+sonorant* |
| | *+continuant* | *+syllabic* | *−syllabic* | *+syllabic* | *−syllabic* |
| | *+strident* | *−back* | *−nasal* | *+back* | *−nasal* |
| | *+voiced* | *+high* | *+rhotic* | *−high* | *+labial* |
| | *+anterior* | *+lax* | | *+low* | |
| $B\Rightarrow$ | Fr | V | SC | V | SC |
| $L\Rightarrow$ | $l_1$ | $l_2$ | $l_3$ | $l_4$ | $l_5$ |
| | Fon | VOP | Son | VOP | Son |
| | Foff | P | D | P | D |
| | | | Soff | | Soff |

bias term $b$ are found from the training data using quadratic optimization methods. Two commonly used kernels are the radial basis function (RBF) kernel and the linear kernel. For the RBF kernel, $K(\mathbf{x}_i,\mathbf{x})=\exp(-\gamma|\mathbf{x}_i-\mathbf{x}|^2)$ where the parameter $\gamma$ is usually chosen empirically by cross validation from the training data. For the linear kernel, $K(\mathbf{x}_i,\mathbf{x})=\mathbf{x}_i.\mathbf{x}+1$.

### B. Probabilistic framework

The problem of speech recognition can be expressed as the maximization of the posterior probability of sets of phonetic features where each set represents a sound or a phoneme. A set of phonetic features include (1) manner phonetic features represented by landmarks and (2) place or voicing phonetic features found using the landmarks. Mathematically, given an acoustic observation sequence $O$, the problem can be expressed as

$$\hat{U}\hat{L} = \arg\max_{U,L} P(U,L|O) = \arg\max_{U,L} P(L|O)P(U|O,L), \tag{2}$$

where $L=\{l_i\}_{i=1}^{M}$ is a sequence of landmarks and $U=\{u_i\}_{i=1}^{N}$ is the sequence bundles of features corresponding to a phoneme sequence. The meaning of these symbols is illustrated in Table II for the digit "zero." Computation of $P(L|O)$ is the process of probabilistic detection of acoustic landmarks given the acoustic observations and the computation of $P(U|L,O)$ is the process of using the landmarks and acoustic observations to make probabilistic decisions on place and voicing phonetic features. The goal of the rest of the paper is to show how $P(L|O)$ is computed for a given landmark sequence and how different landmark sequences and their probabilities can be found given an observation sequence.

There are several points to note with regards to the notation in Table II.

1. $l_i$ denotes a set of related landmarks that occur during the same broad class. For example, the syllabic peak (P) and the vowel onset point (VOP) occur during a vowel. The VOP should occur at the start of the vowel and P should occur during the vowel when the vocal tract is most open. Also, certain landmarks may be redundant in the sequence. For example, when a vowel follows a sonorant consonant, the sonorant consonant offset and the vowel onset are identical.
2. Each set of landmarks $l_i$, as shown in Table III, is related to a broad class $B_i$ of speech selected from the set: vowel (V), fricative (Fr), sonorant consonant (SC), stop burst (ST), silence (SILEN). For example, P and VOP are related to the broad class V. Let $B=\{B_i\}_{i=1}^{M}$ denote the sequence of broad classes corresponding to the sequence of sets of landmarks $L$. Note that, in this paper, that ST denotes the burst region of a stop consonant, and the closure region is assigned the broad class SILEN.
3. The number of broad classes $M$ and the number of bundles of phonetic features $N$ may not be the same in general. This difference may occur because a sequence of sets of landmarks and the corresponding broad class sequence may correspond to one set of phonetic features or two sets of phonetic features. For example, SILEN-ST could be the closure and release of one stop consonant, or it could be that the closure corresponds to one stop consonant and the release corresponds to another stop consonant (e.g., the cluster /kt/ in the word "vector"). Likewise, one set of landmarks or the corresponding broad class may correspond to two sets of place features. For example, in the word "omni" with the broad class sequence V-SC-V, the SC will have the features of the sound /m/ (calculated using the SC onset) as well as the sound /n/ (calculated using SC offset).

The landmarks and the sequence of broad classes can be obtained deterministically from each other. For example, the sequence $B=\{\text{SILEN},\text{Fr},\text{V},\text{SC},\text{V},\text{SC},\text{SILEN}\}$ for "zero" in Table II will correspond to the sequence of sets of landmarks $L$ shown. Therefore

TABLE III. Landmarks and corresponding broad classes.

| Broad class segment | Landmark type |
|---|---|
| Vowel (V) | Syllabic peak (P) |
| | Vowel onset point (P) |
| Stop (ST) | Burst |
| Sonorant consonant (SC) | Syllabic dip (D) |
| | SC onset (Son) |
| | SC offset (Soff) |
| Fricative (Fr) | Fricative onset (Fon) |
| | Fricative offset (Foff) |

$$P(L|O) = P(B(L)|O), \tag{3}$$

where $B(L)$ is a sequence of broad classes for which the landmark sequence $L$ is obtained. Note that the symbols $B$, $U$ and $L$ contain information about the order in which the broad classes or landmarks occur, but they do not contain information about the exact start and end times of each of those units. The equivalence of broad classes and landmarks is not intended as a general statement and it is assumed to hold only for the landmarks and broad classes shown in Table III.

### C. Segmentation using manner phonetic features

Given a sequence of $T$ frames $O=\{o_1, o_2, \ldots, o_T\}$, where $o_t$ is the vector of APs at time $t$ ($t$ is in the units of frame numbers), the most probable sequence of broad classes $B=\{B_i\}_{i=1}^{M}$ and their durations $D=\{D_i\}_{i=1}^{M}$ have to be found. The frame $o_t$ is considered as the set of all APs computed at frame $t$ to develop the probabilistic framework, although EBS does not use all of the APs in each frame. The probability $P(B|O)$ can be expressed as

$$P(B|O) = \sum_{D} P(B,D|O). \tag{4}$$

The computation of $P(B,D|O)$ for a particular $B$ and all $D$ is a very computationally intensive task in terms of storage and computation time. Therefore, an approximation is made that is similar to the approximation made by Viterbi decoding in HMM-based recognition systems and the SUMMIT system (Glass *et al.*, 1996),

$$P(B|O) \approx \max_{D} P(B,D|O). \tag{5}$$

Because the probabilities $P(B|O)$ calculated this way for different $B$ will not add up to one, the more correct approximation is

$$P(B|O) \approx \frac{\max_D P(B,D|O)}{\sum_B \max_D P(B,D|O)}. \tag{6}$$

Provided that a frame at time $t$ lies in the region of one of the manner classes, the posterior probability of the frame being part of a vowel at time $t$ can be written as (see Fig. 2)

$$P_t(V|O) = P_t(+\,speech, +\,sonorant, +\,syllabic|O)$$
$$= P_t(+\,speech|O)P_t(+\,sonorant| + speech, O), \tag{7}$$

$$P_t(+\,syllabic| + sonorant, +\,speech, O), \tag{8}$$

where $P_t$ is used to denote the posterior probability of a feature or a set of features at time $t$. A similar expression can be written for each of the other manner classes.

Calculation of the posterior probability for each feature requires only the acoustic correlates of that feature. Furthermore, to calculate the posterior probability of a manner phonetic feature at time $t$, only the acoustic correlates of the feature in a set of frames $\{t-s, t-s+1, \ldots, t+e\}$, using $s$ previous frames and $e$ following frames along with the current frame $t$, are required to be used. Let this set of acoustic correlates extracted from the analysis frame and the adjoining frames for a feature $f$ be denoted by $x_t^f$. Then Eq. (8) can be rewritten as

$$P_t(V|O) = P_t(+\,speech|x_t^{speech})P_t(+\,sonorant| + speech, x_t^{sonorant})$$
$$P_t(+\,syllabic| + sonorant, +\,speech, x_t^{syllabic}). \tag{9}$$

The probability $P(B,D|O)$ can now be expanded in terms of the underlying manner phonetic features of each broad class. Denote the features for class $B_i$ as the set $\{f_1^i, f_2^i, \ldots, f_{N_{B_i}}^i\}$, the broad class at time $t$ as $b_t$, and the sequence $\{b_1, b_2, \ldots, b_t\}$ as $b^t$. Note that $B$ is the broad class sequence with no information about the duration of each broad class in the sequence. On the other hand, $b_t$ denotes a broad class at frame $t$. Therefore, the sequence $b^t$ includes the information of durations of each of the broad classes until time $t$. Using this notation, the posterior probability of a broad class sequence $B$ and durations $D$ can be expanded as

$$P(B,D|O) = \prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} P_t(B_i|O, b^{t-1}). \tag{10}$$

The variable $t$ in the above equation is the frame number, the limits of which can be explained as follows. $\Sigma_{j=1}^{i-1}D_j$ is the sum of the durations of the $i-1$ broad classes before the broad class $i$, and $\Sigma_{j=1}^{i}D_j$ is the sum of durations of the first $i$ broad classes. Then, $\Sigma_{j=1}^{i}D_j - \Sigma_{j=1}^{i-1}D_j$ is the duration of the $i$th broad class. Therefore, numbers $\{1+\Sigma_{j=1}^{i-1}D_j, 2+\Sigma_{j=1}^{i-1}D_j, \ldots, D_i+\Sigma_{j=1}^{i-1}D_j\}$ are the frame numbers of the frames that occupy the $i$th broad class. When the lower and upper limits of $t$ are specified as $1+\Sigma_{j=1}^{i-1}D_j$ and $D_i+\Sigma_{j=1}^{i-1}D_j$, respectively, it means that the product is taken over all the frames of the $i$th broad class.

Making a stronger use of the definition of acoustic correlates by assuming that the acoustic correlates of a manner feature at time $t$ are sufficient even if $b^{t-1}$ is given,

$$P(B,D|O) = \prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|x_t^{f_{k^i}}, f_1^i, \ldots, f_{k-1}^i, b^{t-1}). \tag{11}$$

Now expanding the conditional probability,

$$P(B,D|O) = \prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i, x_t^{f_k^i}, f_1^i, \ldots, f_{k-1}^i, b^{t-1})}{P_t(x_t^{f_k^i}, f_1^i, \ldots, f_{k-1}^i, b^{t-1})}. \tag{12}$$

Splitting the priors gives

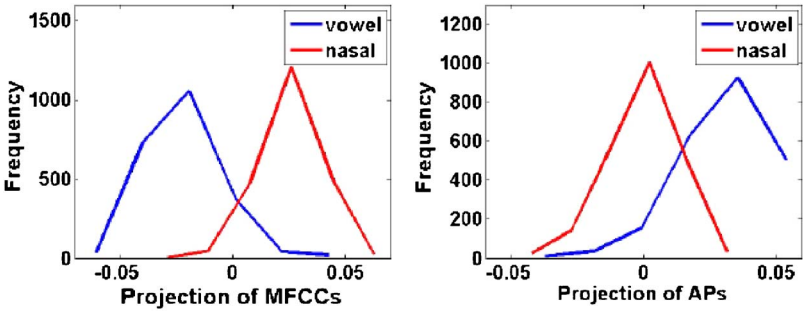


FIG. 5. (Color online) (a) Projection of 39 MFCCs into a one-dimensional space with vowels and nasals as discriminating classes, (b) similar projection for four APs used to distinguish +sonorant sounds from −sonorant sounds. Because APs for the sonorant feature discriminate vowels and nasals worse than MFCCs, they are more invariant.

$$P(B,D|O) = \prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|f_1^i,\ldots,f_{k-1}^i,b^{t-1}) \times \frac{P_t(x_t^{f_k^i}|f_1^i,\ldots,f_k^i,b^{t-1})}{P_t(x_t^{f_k^i}|f_1^i,\ldots,f_{k-1}^i,b^{t-1})}. \quad (13)$$

The probability terms not involving the feature vector $x_t^{f_k^i}$ can now be combined to get the prior probabilities of the broad class sequence and the sequence dependent durations, that is,

$$\prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|f_1^i,\ldots,f_{k-1}^i,b^{t-1}) = P(B,D) = P(B)P(D|B). \quad (14)$$

Now given the set $\{f_1^i,\ldots,f_{k-1}^i\}$ or the set $\{f_1^i,\ldots,f_k^i\}$, $x_t^{f_k^i}$ is assumed to be independent of $b^{t-1}$. This independence of the APs from the previous broad class frames is hard to establish, but it can be shown to hold better for the knowledge-based APs than for the mel-frequency cepstral coefficients (MFCCs) (see Fig. 5) under certain conditions as discussed in Sec. III B. In words, this independence means that given a phonetic feature or the phonetic features above that feature in the hierarchy, the APs for that phonetic feature are assumed to be invariant with the variation of the broad class labels of the preceding frames. For example, the APs for the feature *sonorant* in a *+sonorant* frame are assumed to be invariant of whether the frame lies after vowel, nasal or fricative frames. This is further discussed in Sec. III B. Making this independence or invariance assumption and applying Eq. (14) in Eq. (13),

$$P(B,D|O) = P(B)P(D|B)\prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} \times \prod_{k=1}^{N_{B_i}} \frac{P_t(x_t^{f_k^i}|f_1^i,\ldots,f_k^i)}{P_t(x_t^{f_k^i}|f_1^i,\ldots,f_{k-1}^i)}, \quad (15)$$

which can be rewritten as

$$P(B,D|O) = P(B)P(D|B)\prod_{i=1}^{M} \prod_{t=1+\Sigma_{j=1}^{i-1}D_j}^{D_i+\Sigma_{j=1}^{i-1}D_j} \times \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i|x_t^{f_k^i},f_1^i,\ldots,f_{k-1}^i)}{P_t(f_k^i|f_1^i,\ldots,f_{k-1}^i)}. \quad (16)$$

The posterior $P_t(f_k^i|x_t^{f_k^i},f_1^i,\ldots,f_{k-1}^i)$ is the probability of the binary feature $f_k^i$ obtained using the APs $x_t^{f_k^i}$ and it is obtained in this work from an SVM-based binary classifiers as described in Sec. I below. The term $P_t(f_k^i|x_1^i,\ldots,f_{k-1}^i)$ normalizes the imbalance of the number of positive and negative samples in the training data. The division on the right side of Eq. (16) can be considered as the conversion of a posterior probability to a likelihood by division by a prior. The prior is computed as the division of the number of training samples for the positive value of the feature to the number of training samples for the negative value of the feature.

### 1. Training and application of binary classifiers

One SVM classifier was trained for each of the phonetic features shown in Fig. 2. The input to the classifier is the set of APs shown in Table I for that feature. The sounds used to get the training samples of class +1 and class −1 for each SVM are shown in Table IV. For the feature *continuant*, the

TABLE IV. Sounds used in training of each classifier.

| Phonetic feature | Sounds with +1 label | Sounds with −1 label |
|---|---|---|
| *speech* | All speech sounds excluding stop closures | Silence, pauses and stop closures |
| *sonorant* | Vowels, nasals and semivowels | Fricatives, affricates and stop bursts |
| *syllabic* | Vowels | Nasals and semivowels |
| *continuant* | Fricatives | Stop bursts |

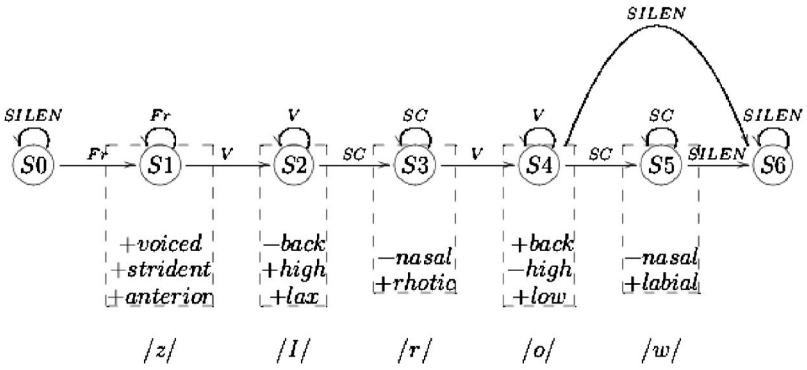


FIG. 6. A phonetic feature-based pronunciation model for the word "zero."

stop burst frame identified as the first frame of a stop consonant using TIMIT labeling was used to extract APs representative of the value −1 of that feature. For the +1 class of the feature *continuant*, the APs were extracted from all of the fricative frames. For the other features, all frames for each of the classes were extracted as training samples. Flap (/dx/), syllabic sonorant consonants (/em/, /el/, /en/, /er/ and /eng/) and diphthongs (/iy/, /ey/, /ow/, /ay/, /aw/, and /uw/) were not used in the training of the feature *syllabic*, and affricates (/jh/ and /ch/) and glottal stops were not used in training of the feature *continuant*, but these sounds were used for frame-based testing. The reason for not using these sounds for training is that they have different manifestations. For example, the affricates /ch/ and /jh/ may appear with or without a clear stop burst. However, such information is not marked in the TIMIT hand-transcribed labels.

SVM Light (Joachims, 1998), an open-source toolkit for SVM training and testing, was used for building and testing the classifiers. Two types of SVM kernels were used—linear and radial basis function (RBF)—to the build corresponding two types of classifiers. The optimal number of adjoining frames $s$ and $e$ used in each classifier as well as the optimal SVM related parameters (e.g., the bound on slack variables $\alpha_i$ and $\gamma$ for RBF kernels) were found using cross validation over a separate randomly chosen data set from the TIMIT training set.

A SVM outputs a real number for a teat sample. To convert this real number into a probability, the real space of the SVM outputs is divided into 30 bins of equal sizes between −3 and +3. This range was chosen empirically from observations of many SVM outputs. After the SVMs are trained, the proportion of samples of class +1 to the total numbers of training samples in each of the bins is noted and the proportions for all of the bins are stored in a table. While testing, the bin corresponding to the real number obtained for a particular test sample is noted and its probability is looked up from the stored table.

#### 2. Probabilistic segmentation

A Viterbi-like probabilistic segmentation algorithm (Juneja, 2004) takes as input the probabilities of the manner phonetic features—*sonorant*, *syllabic*, *continuant*—and silence from the SVM classifiers and outputs the probabilities $P(B|O)$ under the assumption of Eq. (5). The algorithm is similar to the one used by Lee (1998). The algorithm operates on the ratio of posterior probabilities on the right side of Eq. (16), unlike the algorithm developed by Lee (1998) where segment scores of observations in speech segments are used. Another difference is that the transition points in the segmentation algorithm in the current work are obtained as those frames at which the ranking of the posterior probabilities of the broad classes changes. In the work by Lee (1998), the transitions points were calculated from the points of significant change in the acoustic representation.

### D. Deterministic location of landmarks

Once a broad class sequence with the start and end times of each of the broad classes is found, the landmarks are located deterministically. Fon and Foff are allotted the start and end times of the broad class Fr. Son and Soff are assigned the start and end times of the broad class SC. The stop burst B is found as the location of the maximum value of the temporal onset measure within a 60 ms window centered at the first frame of the segment ST. VOP is assigned the first frame of the segment V, and P is assigned the location of highest value of $E[640,2800]$ in the segment V. The syllabic dip D for an intervocalic SC is assigned the location of the minimum in $E[640,2800]$ in the segment SC. For prevocalic and postvocalic SC, D is assigned the middle frame of the SC segment.

### E. Constrained landmark detection for word recognition

For isolated word or connected word recognition, manner class segmentation paths are constrained by a pronunciation model in the form of a finite state automata (FSA) (Jurafsky and Martin, 2000). Figure 6 shows an FSA-based pronunciation model for the digit "zero" and the canonical pronunciation /z I r ow/. The broad manner class representation corresponding to the canonical representation is Fr-V-SC-V-SC (the last SC is for the off glide of /ow/). The possibility that the off-glide of the final vowel /ow/ may or may not be recognized as a sonorant consonant is represented by a possible direct transition from the V state to the SILEN state. Starting with the start state S0, the posterior probability of a particular path through the FSA for zero can be calculated using the likelihood of a transition along a particular broad class $B_i$ as

$$\prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i|x_t^{f_k^i}, f_1^i, \ldots, f_{k-1}^i)}{P_t(f_k^i|f_1^i, \ldots, f_{k-1}^i)}.$$

The likelihoods of all of the state transitions along a path are multiplied with the prior $P(B)$ and the duration densities $P(D|B)$ using the durations along that path. The probabilistic segmentation algorithm gives for an FSA and an observation sequence the best path and the posterior probability computed for that path. Note that word posterior probabilities can be found by multiplying the posterior probability $P(L|O)$ of the landmark sequence with the probability $P(U|OL)$ of the place and voicing features computed at the landmarks (Juneja, 2004), about computing complete word probabilities is out of the scope of this paper.

## III. EXPERIMENTS AND RESULTS

### A. Database

The "si" and "sx" sentences from the training section of the TIMIT database were used for training and development. For training the SVM classifiers, randomly selected speech frames were used because SVM training with all of the available frames was impractical. For training the HMMs, all of the si and sx sentences from the training set were used. All of the si and sx sentences from the testing section of the TIMIT database were used for testing how well the systems perform broad class recognition. The 2240 isolated digit utterances from the TIDIGITS training corpus were used to obtain word-level recognition results. If spoken canonically, the digits are uniquely specified by their broad class sequence. Thus, word-level results are possible for this constrained database. Note that the TIMIT database is still used for training since the TIDIGITS database is not transcribed. Thus, this experiment not only shows how well the systems perform word-level recognition, but it also allows for cross-database testing.

### B. Sufficiency and invariance

In this section, an illustration of how the APs satisfy the assumptions of the probabilistic framework better than the MFCCs is presented. Although it is not clear how sufficiency and invariance can be rigorously established for certain parameters, some idea can be obtained from classification and scatter plot experiments. For example, sufficiency of the four APs used for the sonorant feature detection—$E[0,F3]$, $E[100\ \text{Hz},400\ \text{Hz}]$, $E[F3,f_s/2]$, ratio of the $E(0,F3)$ to the energy in ($F3$, half of sampling rate)[1]—can be viewed in relation to the 39 mel-frequency cepstral coefficients (MFCCs) in terms of classification accuracy of the sonorant feature. Two SVMs with linear kernels were trained, one for the APs and one for the MFCCs, using a set of 20,000 randomly selected sample frames of each of the *+sonorant* and *−sonorant* frames from dialect region 1 of the TIMIT training set. The same number of samples were extracted from dialect region 8 of the TIMIT training set for testing. A frame classification accuracy of 93.0% was obtained on data using the APs and SVMs, which compares well to 94.2% accuracy obtained using the MFCCs and SVMs. Note that for the two SVMs the same speech frames were used for training as well as testing and only the types of acoustic features were different.

In Eq. (16), the APs $x_t^{f_k^i}$ for a manner feature were assumed to be independent of the manner class labels of the preceding frames $b^{t-1}$ when either $\{f_1^i,\ldots,f_k^i\}$ or $\{f_1^i,\ldots,f_{k-1}^i\}$ was given. For example, for the feature $f_k^i=+sonorant$, the set $\{f_1^i,\ldots,f_k^i\}$ is $\{+speech,+sonorant\}$ and $\{f_1^i,\ldots,f_{k-1}^i\}$ is $\{+speech\}$. Consider the case where $\{f_1^i,\ldots,f_k^i\}$ is given, that is, the value of the feature whose APs are being investigated is known. A typical case where the assumption may be hard to satisfy is when the APs for the *sonorant* feature are assumed to be invariant of whether the analysis frame lies in the middle of a vowel region or the middle of a nasal region (both vowels and nasals are *+sonorant*).That is, $b^{t-1}$ will be composed of nasal frames in one case and vowel frames in the other case.

Such independence can roughly be measured by the similarity in the distribution of the vowels and nasals based on the APs for the feature *sonorant*. To test this independence, *sonorant* APs were extracted from dialect region 8 of the TIMIT training set from each of the nasal and vowel segments. Each set of APs was extracted from a single frame located at the center of the vowel or the nasal. The APs were then used to discriminate vowels and nasals using Fischer linear discriminant analysis (LDA). Figure 5(a) shows the distribution of the projection of the 39 MFCCs extracted from the same 200 frames into a one-dimensional space using LDA. A similar projection is shown for the four *sonorant* APs in Fig. 5(b). It can be seen from these figures that there is considerably more overlap in the distribution of the vowels and the nasals for the APs of the *sonorant* feature than for the MFCCs. Thus, the APs for the sonorant feature are more independent of the manner context than are the MFCCs. The overlap does not show worse performance of the APs compared to MFCCs because the *sonorant* APs are not meant to separate vowels and nasals. They separate vowels, nasals and semivowels (i.e., sonorants) from fricatives, stop consonants and affricates (i.e., obstruents). Thus, the APs for the feature *sonorant* are invariant across different *+sonorant* sounds but successfully discriminate *+sonorant* sounds from *−sonorant* sounds. Further discussion of the sufficiency and the invariance properties of the APs can be found in Juneja (2004).

### C. Frame-based results

The SVMs for each feature utilized APs extracted from the analysis frame as well as $s$ starting frames and $e$ ending frames. The values of the two variables $e$ and $s$ were obtained for each classifier by performing validation over a subset of the TIMIT training data (Juneja, 2004). Training was performed on randomly picked samples (20,000 samples for each class) from the si and sx sentences of the TIMIT training set. The binary classification results on the whole of the TIMIT test set at the optimal values of $s$ and $e$ are shown in Table V in two cases—(1) when all the frames were used for testing and (2) when only the middle one-third portion of each broad class was used for testing. The difference in the results indicates the percentage of errors that are made due to boundary or coarticulation effects. Note that in the presented landmark-based system, it is not important to classify each frame correctly. The results on the middle one-third segment

TABLE V. Binary classification results for manner features in %. Accuracy on middle frames is not shown for the feature *continuant* because the feature distinguishes the stop releases from the beginning of fricatives.

| Feature | *s* | *e* | Accuracy on middle frames | Accuracy on all frames |
|---|---|---|---|---|
| *sonorant* | 4 | 1 | 96.59 | 94.39 |
| *syllabic* | 16 | 24 | 85.00 | 80.06 |
| Speech/silence | 3 | 2 | 94.60 | 93.50 |
| *continuant* | 4 | 4 | … | 95.58 |

are more representative of the performance of the system because if the frames in a stable region are correctly recognized for a particular manner feature, this would mean that the corresponding landmarks may still be correctly obtained. For example, if the middle frames of an intervocalic sonorant consonant are correctly recognized as *syllabic*, then the correct recognition of frames near the boundary is not significant because landmarks for the sonorant consonant will be obtained accurately. For the feature *continuant*, the classification error on middle frames is not relevant because the SVM is trained to extract the stop burst as opposed to a certain stable region of speech. Also the transient effects at broad class boundaries are minimized by low probability density values of very small broad class durations.

Figures 7–10 show the most significant sources of error for each of the phonetic features. The errors include misclassifications of the +*feature* sounds as −*feature*, and vice versa. For the feature *sonorant*, it can be seen that the sounds /v/ and the glottal stop /q/ are often detected as +*sonorant*. The sound /v/ is often manifested as a sonorant consonant so that the assignment of +*sonorant* for /v/ is expected. In the case of the glottal stop, a separate detector is required either at the broad class recognition level or further down the hierarchy to recognize glottalization because it can be significant for lexical access, especially in the detection of the consonant /t/ (Stevens, 2002). For the feature *syllabic*, classification accuracy for nasals as -syllabic is above 90%. But the semivowels—/ɹ/, /r/, /l/ and /w/ have lower accuracies which is expected because of the vowel-like behavior of these sounds. About 30% of the frames of reduced vowels are also misrecognized as sonorant consonants. This typically happened when a sonorant consonant followed a stressed vowel and preceded a reduced vowel such that the reduced vowel is confused as a continuation of the sonorant consonant. A similar result was shown by Howitt (2000) where the vowel landmarks were missed for reduced vowels more than other vowels. The performance of the feature *continuant* is 95.6% which indicates the accuracy on classification of onset frames of all nonsonorant sounds. That is, an error was counted if a stop burst was wrongly classified as −*continuant* or a fricative onset was wrongly classified as a stop burst. The major source of error is the misclassification of 13.7% of fricative onsets as stop bursts.

### D. Sequence-based results

The SVM models obtained in the frame-based analysis procedure were used to obtain broad class segmentation as well as the corresponding landmark sequences for all of the si and sx sentences of the TIMIT test set using the probabilistic segmentation algorithm. Not all broad class sequences were allowed as the segmentation paths were constrained using a pronunciation graph such that (1) SCs only occur adjacent to vowels, (2) ST is always preceded by SILEN (for stop closure) and (3) each segmentation path starts and ends with silence. The same pronunciation graph was used for both the EBS system and the HMM system. The duration probability for each broad class was modeled by a mixture of Rayleighs using a single Rayleigh density for the classes SC,

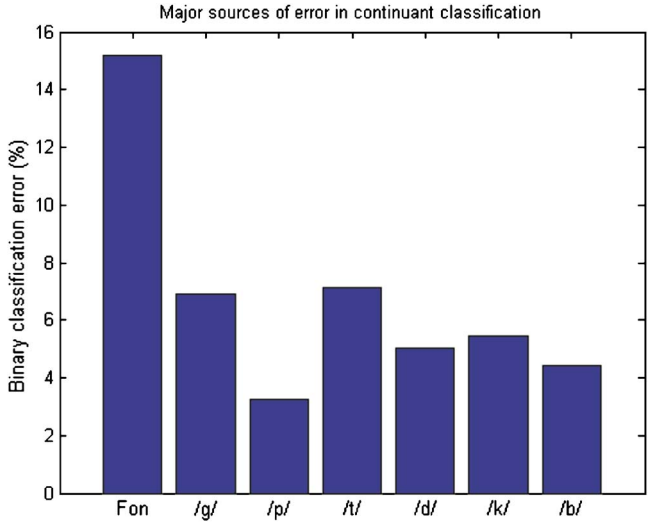


FIG. 7. (Color online) Sounds with high error percentages for the feature *sonorant*; "voic-stop" represents voiced stop consonants.

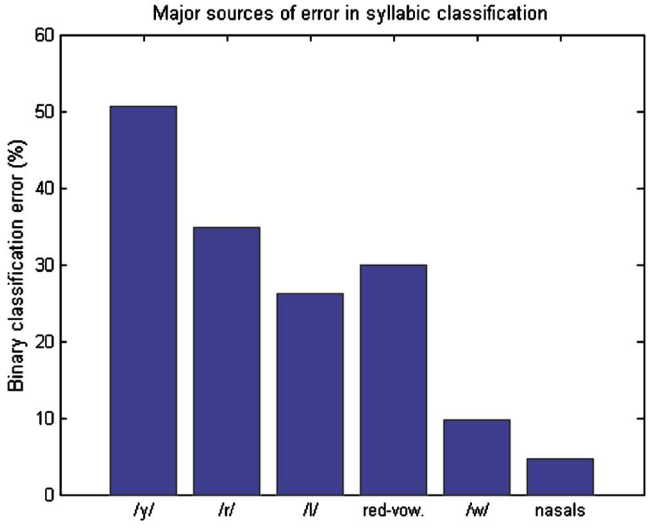


FIG. 8. (Color online) Sounds with high error percentages for the feature *continuant*. Fon represents fricative onsets.

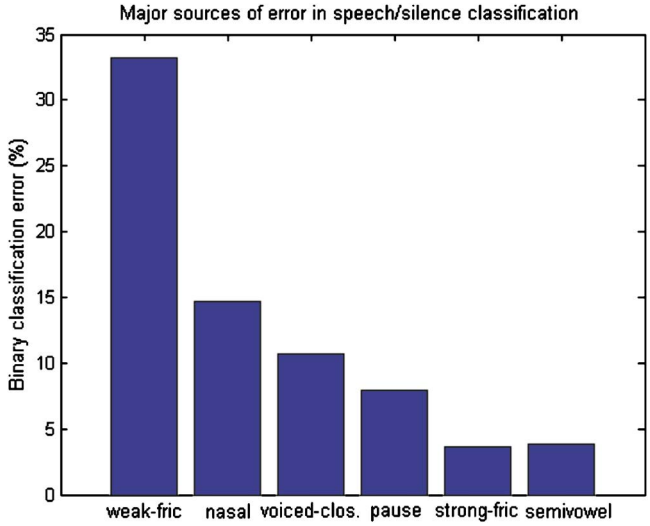


FIG. 9. (Color online) Sounds with high error percentages for the feature *syllabic*. "Red-vow" represents reduced vowels.

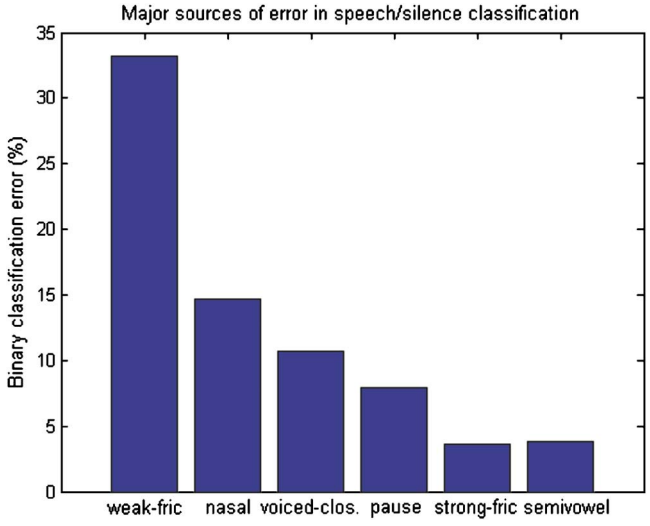


FIG. 10. (Color online) Sounds with high error percentages in speech silence distinction. "Weak-fric" represents weak fricatives and "strong-fric" represents strong fricatives. "Voiced-clos" represents closures of voiced stop consonants.

V, Fr and ST, and a mixture of two Rayleigh densities for SILEN (one density targets short silence regions like pauses and closures and the other density targets beginning and ending silence). The parameter for each Rayleigh density was found using the empirical means of the durations of each of the classes from the TIMIT training data.

For the purpose of scoring, the reference phoneme labels from the TIMIT database were mapped to manner class labels. Some substitutions, splits and merges as shown in Table VI were allowed in the scoring process. Specifically, note that two identical consecutive broad classes were allowed to be merged into one since the distinction between such sounds is left to the place classifiers. Also note that affricates were allowed to be recognized as ST+Fr as well as Fr, and similarly diphthongs—-/iʏ/, /eʏ/, /ow/, /aʏ/, /aw/, and /uw/—were allowed to be recognized as V+SC as well as V because the off glides may or may not be present. Scoring was done on the sequences of hypothesized symbols without using information of the start and end of the broad class segments which is similar to word level scoring in continuous speech recognition.

The same knowledge-based APs were used to construct an 11-parameter front end for a HMM-based broad class segmentation system. The comparison with the HMM-based system does not show that the presented system performs superior or inferior to the HMM-based systems, but it shows an acceptable level of performance. The HMM-based systems have been developed and refined over decades and the work presented in this paper is only the beginning of the development of a full speech recognition system based on phonetic features and acoustic landmarks.

All the HMMs were context-independent three-state (excluding entry and exit states) left-to-right HMMs with diagonal covariance matrices and eight-component mixture observation densities for each state. All the si and sx utterances from the TIMIT training set were used for training the HMM broad classifier. A HMM was built for each of the five broad classes, and a separate HMM was built for each of the special sounds—affricate, diphthong, glottal stop, syllabic sonorant consonant, flap /dx/ and voiced aspiration /hv/—making a total of 11 HMMs. Only the five broad class models were used in testing and the HMMs for the special sounds were ignored, so that the training and testing sounds of HMM and EBS were identical. The HMM models were first initialized and trained using all of the training segments for each model separately (for example, using semivowel and nasal segments for the sonorant consonant model), and then improved using embedded training on the concatenated HMMs for

TABLE VI. Allowed splits, merges and substitutions.

| Reference | Allowed hypothesis | Reference | Allowed hypothesis |
|---|---|---|---|
| V+V | V | SC+SC | SC |
| Fr+Fr | Fr | SILEN+SILEN | SILEN |
| /q/+V, V+/q/ | V | /q/ | ST, SC |
| /t/, /p/, /k/, /g/, /d/ | ST+Fr | /v/ | SC, Fr |
| /em/, /en/, /er/, /el/ | V+SC | /ch/, /jh/ | ST+Fr |
| /hv/ | SC, Fr | /dx/ | SC |
| /dx/ | SILEN+ST | /iʏ/, /ow/, /eʏ/, /oʏ/, /aw/, /uw/, /ow/ | V+SC |

TABLE VII. Broad class segmentation results in percent. Correctness (Corr)/Accuracy (Ace) are shown when the system is scored on the basis of numbers of deletions, insertions and substitutions of broad classes. A "-" in a cell means that the particular system was computationally too intensive to get a result from.

| | EBS (RBF) | EBS (linear) | HMM |
|---|---|---|---|
| | Corr/Acc | Corr/Acc | Corr/Acc |
| 11 APs | 86.2/79.5 | 84.0/77.1 | 80.9/73.7 |
| 39 MFCCs | - | 86.1/78.2 | 86.8/80.0 |

each sentence. Triphone models and other similarly improved HMM models may give better results than the ones presented in this paper, but the focus here is to build a base line HMM system to which EBS's performance can be compared.

The results are shown in Table VII. The results are also shown for EBS for two different front ends—AP and MFCC (including MFCCs, their delta and acceleration coefficients which gives a 39-parameter front end). The performance of all of the systems, except when EBS is used with MFCCs, is comparable although the HMM-MFCC system gives the maximum accuracy. The inferior performance of the MFCCs with EBS is perhaps because of the better agreement of APs with the invariance assumptions of the probabilistic framework. Similarly, better performance of MFCCs in the HMM framework may be because of better agreement with the diagonal covariance assumption of the HMM system applied here. That is, APs are not processed by a diagonalization step prior to application to the HMM systems while MFCCs go through such a process. These are possible explanations of these results and they are open to further investigation.

An example of landmarks generated by EBS on a test sentence of TIMIT is shown in Fig. 11 which also shows how errors in the system can be analyzed. The pattern recognizer calls the /dh/ in "the" (as marked by the ellipse) a sonorant constant (SC) instead of the correct broad class Fr. The reason is that the parameter $E[0,F3]/E[F3,f_s/2]$ does not dip adequately as it usually does in most *–sonorant sounds*. This indicates that improved APs, for example, from the APP detector (Deshmukh *et al.*, 2005) that directly captures the aperiodicity, are needed to correct errors like this one.

The confusion matrix of various landmarks for EBS using the AP front end is shown in Table VIII without including the sounds—diphthongs, syllabic sonorant consonants, flaps, /v/, affricates and the glottal stop /q/. For this latter set of sounds the confusion matrix is shown in Table IX. There is a considerable number of insertion errors. Insertions are common in any speech recognition system because typical speaking rates vary from training segments to test segments. There are sudden onsets of vowels and fricatives that give rise to stop burst insertions; 68% of stop burst insertions were at the begining of fricative segments and 46% were at the beginning of the sentences possibly because speakers are highly likely to start speaking with a sudden onset. High-frequency noise in the silence regions and aspiration at the end or beginning of vowels cause fricative insertions; 44% of all fricative insertions occur with an adjoining silence region, 43% of the rest of the fricative insertions have an adjoining vowel.

### E. Word-level results and constrained segmentation results

The SVM and HMM models obtained by training on the TIMIT database were then applied to the isolated digits of the TIDIGITS database in both the vocabulary constrained

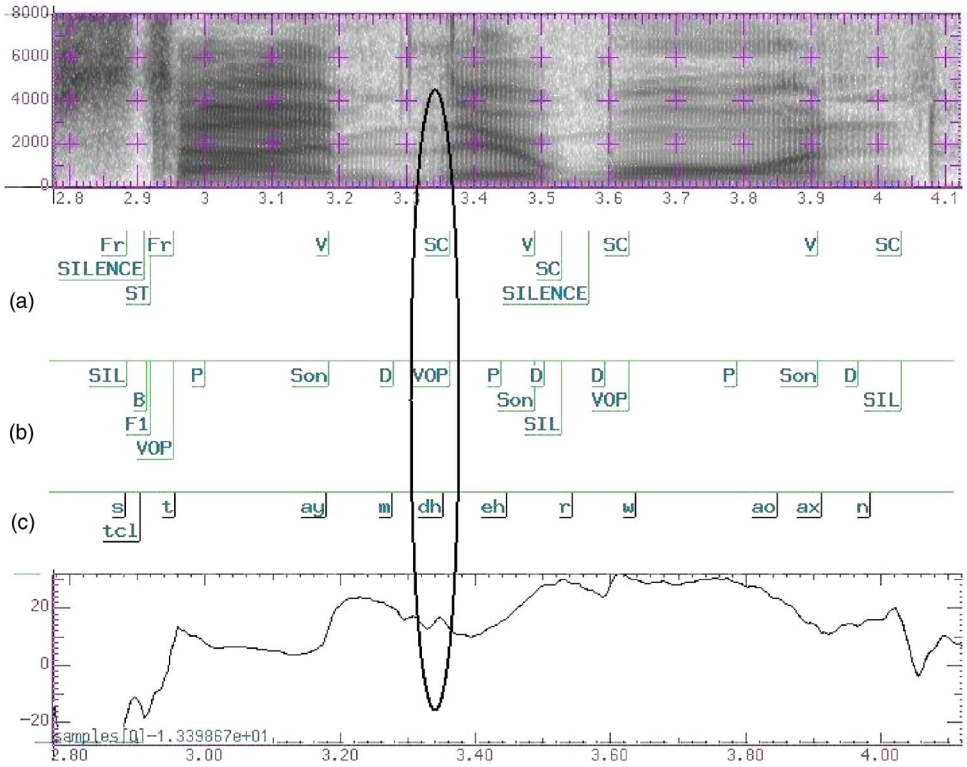


FIG. 11. (Color online) Top: spectrogram of the utterance, "time they're worn." A: Broad class labels, B: Landmark labels, C: phoneme labels, bottom: ratio of $E[0,F3]$ to $E[F3,f_s/2]$. Broad class and phoneme labels are marked at the end of each sound, and the landmark labels show the time instant of each landmark. The ellipse shows an error made by the system on this utterance. $E[0,F3]/E[F3,f_s/2]$ does not dip in the /dh/ region which makes the pattern recognizer call the fricative a *+sonorant* sound.

TABLE VIII. Confusion matrix for landmarks with exclusion of affricates, syllabic sonorant consonants, /v/, glottal stop /q/, diphthongs and flap /dx/. Only nonredundant landmarks are shown. For example, VOP implies presence of a syllabic peak P and vice versa, therefore, only VOP is used in the confusion matrix.

| | Total | Fon | SIL | VOP | Son | B | Deletions | Correct (%) |
|---|---|---|---|---|---|---|---|---|
| Fon | 6369 | 5607 | 10 | 1 | 136 | 185 | 430 | 88.03 |
| SIL | 10,232 | 15 | 9281 | 12 | 104 | 0 | 820 | 90.71 |
| VOP | 12,467 | 50 | 56 | 11,146 | 18 | 24 | 1173 | 89.40 |
| Son | 5504 | 155 | 65 | 1 | 4565 | 95 | 1290 | 70.82 |
| B | 9152 | 448 | 2 | 24 | 104 | 2755 | 797 | 84.98 |
| Insertions | 3439 | 682 | 692 | 206 | 1038 | 821 | | |

and the unconstrained modes. In the unconstrained mode, the models were tested in exactly the same way as on the TIMIT database. To get the results on constrained segmentation, the segmentation paths were constrained using the broad class pronunciation models for the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. The segmentation was identically constrained for both the HMM system and EBS. The results are shown in Table X for EBS (with linear as well as RBF kernels) and for the HMM systems trained on TIMIT and tested on TIDIGITS. On moving from unconstrained to constrained segmentation, a similar improvement in performance of the EBS (RBF) and HMM-AP systems can be seen in this table. This result shows that EBS can be constrained in a successful manner as for the HMM system. The overall performance of EBS using RBFs is also very close to the HMM-AP system. HMM-AP system shows better generalization than the HMM-MFCC system over cross-database testing which may be attributed to better speaker independence of the APs compared to the MFCCs (Deshmukh *et al.*, 2002).

Figure 12 shows an example of the output of the unconstrained probabilistic segmentation algorithm for the utterance "two" with canonical pronunciation /t uw/. The two most probable landmark sequences obtained from the algorithm are shown in this figure. The landmark sequence obtained with the second highest probability for this case is the correct sequence. It is hoped that once probabilistic place and voicing decisions are made, the second most probable sequence of landmarks will yield an overall higher posterior word probability for the word two.

Finally, word level accuracies were obtained for all of the systems. The state-of-the-art word recognition accuracy using word HMM models on TIDIGITS is above 98% (Hirrsh and Pearce, 2000). Recognition rates of 99.88% have also been obtained when using word HMM models for recognition of the TI-46 isolated digit database (Deshmukh *et al.*, 2002). These results were obtained using the same topology as in the present experiments (i.e., three-state HMMs with eight-mixture components). The difference is that instead of three-state HMM word models, we are now using three-state HMM broad class models to make the comparison with EBS. Note that a full word recognition system including place features is not presented here and only broad class models are presented. Therefore, a complete segmentation for a digit was scored as correct if it was an acceptable broad class segmentation for that digit.

The results are shown in Table XI. A fully correct segmentation of 68.7% was obtained using the EBS-AP system. About 84.0% of the digits had a correct segmentation among the top two choices. Note that the top two or three choices can be combined with place information to get final probabilities of words. A significant increase in correct recognition in the top two choices over the top one choice shows that there is a good scope of recovery of errors when place information is added. An accuracy of 67.6% was obtained by the HMM-AP system and an accuracy of 63.8% was obtained by the HMM-MFCC system. These results further confirm the comparable performance of the EBS and the HMM-AP systems. Specifically this result shows that a system that selectively uses knowledge based APs for phonetic feature detection can be constrained as well as the HMM systems for limited vocabulary tasks and can also give a similar performance in terms of recognition accuracy.

## IV. DISCUSSION

A landmark-based ASR system has been described for generating multiple landmark sequences of a speech utterance along with a probability of each sequence. The land-

TABLE IX. Confusion matrix for affricates, syllabic sonorant consonants (SSCs), /v/, glottal stop /q/, diphthongs and flap /dx/. Empty cells indicate that those confusions were scored as correct but the exact number of those confusions were not available from the scoring program.

| | Total | Fon | SIL | VOP | Son | B | Deletions | Correct (%) |
|---|---|---|---|---|---|---|---|---|
| /q/ | 927 | 2 | | 0 | 5 | | | 99.25 |
| Diph | 4390 | 23 | 18 | 3991 | 25 | 5 | 328 | 90.91 |
| SSCs | 1239 | 11 | 14 | 1071 | 27 | 1 | 115 | 86.44 |
| /v/ | 710 | | 40 | 2 | | 4 | 198 | 65.63 |
| /dx/ | 632 | 40 | 6 | 0 | | | 75 | 80.85 |
| /ch/, /jh/ | 570 | 562 | 0 | 1 | 0 | | 7 | 98.60 |
| /hv/ | 233 | | 0 | 0 | | 3 | 43 | 80.26 |

TABLE X. Broad class results on TIDIGITS (Correct/Accurate in percent).

| | EBS (linear) | EBS(RBF) | HMM-MFCC | HMM-AP |
|---|---|---|---|---|
| Constrained | 91.7/82.8 | 92.6/85.2 | 92.4/84.3 | 92.3/85.8 |
| Unconstrained | 89.5/64.0 | 93.0/74.3 | 88.6/74.1 | 84.2/72.9 |

mark sequences can be constrained using broad class pronunciation models. For unconstrained segmentation on TIMIT, an accuracy of 79.5% is obtained assuming certain allowable splits, merges and substitutions that may not affect the final lexical access. On cross database constrained detection of landmarks, a correct segmentation was obtained for about 68.7% of the words. This compares well with a correct segmentation for about 67.6% of the words for the HMM system using APs and 63.8% for the HMM system using MFCCs. The percentage accuracy of broad class recognition improved from 74.3% for unconstrained segmentation to 84.2% for constrained segmentation which is very similar to the improvement from 72.9 to 85.5% for the HMM system using APs. These results show that EBS can be constrained by a higher level pronunciation model similar to the HMM systems.

The comparison with previous work on phonetic feature detection is very difficult because of the different test conditions, definitions of features and levels of implementation used by different researchers. At the frame level, the 94.4% binary classification accuracy on the *sonorant* feature compares well with previous work by Bitar (1997) where an accuracy of 94.6% for sonorancy detection on the same database was obtained. The *continuant* result of 95.6% is not directly comparable with previously obtained stop detection results (Bitar, 1997; Liu, 1996; Niyogi, 1998). In the work by Niyogi (1998) results were presented at a frame rate of 1 ms, and in the work by Liu (1996) and Bitar (1997), results were not presented at the frame level. A full probabilistic landmark detection system was not developed in the research cited above. An 81.7% accuracy on the *syllabic* feature may seem low, but note that there is usually no sharp boundary between vowels and semivowels. Therefore, a very high accuracy at the frame level for this feature is not only very difficult to achieve, but also it is not very important as long as sonorant consonants are correctly detected. The authors have not been able to find a previous result to which this number can be suitably compared. At the sequence level, the overall accuracy of 79.5% is comparable to 77.8% accuracy obtained in a nonprobabilistic version of EBS (Bitar, 1997). Note that the most significant improvement over the system by Bitar (1997) is that the current system can be constrained for limited vocabulary and it can be used for obtaining multiple landmark sequences instead of one. There are various other systems to which segmentation results can be compared (Salomon *et al.*, 2004; Ali, 1999), but the comparison is omitted in this work because the purpose of this paper is to present how the ideas from such systems can be applied to a practical speech recognizer.

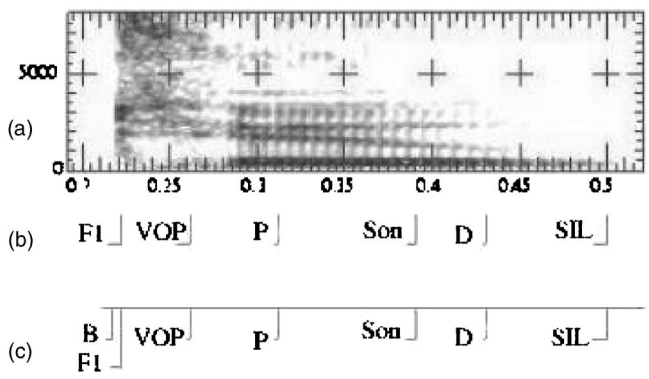


FIG. 12. A sample output of the probabilistic landmark detection for the digit "two." The spectrogram is shown in (a).Two most probable landmark sequences (b) and (c) are obtained by the probabilistic segmentation algorithm. The first most probable sequence (b) has a missed stop consonant but the second most probable sequence gets it.

The complete system for word recognition is currently being developed. There has been some success in small vocabulary isolated word recognition (Juneja, 2004) and in landmark detection for large vocabulary continuous speech recognition (Hasegawa-Johnson *et al.*, 2005). EBS benefits directly from research in discriminative APs for phonetic features, therefore, the system will improve as more powerful APs are designed for various phonetic features. By the use of APs specific to each phonetic feature EBS provides a platform for the evaluation of new knowledge gained on discrimination of different speech sounds. EBS provides easier evaluation of newly designed APs than HMM based systems. If certain APs give better performance in binary classification of phonetic features and are more context independent than currently used APs, then they will give overall better recognition rates. Therefore, complete speech recognition experiments are not required in the process of designing the APs. In the future, apart from the research that will be carried out on the automatic extraction of APs for all the phonetic features, further research will be done on better glide detection and incorporation of previous research (Howitt, 2000) on detection of vowel landmarks. APs to separate nasals from semivowels (Pruthi and Epsy-Wilson, 2003) and to detection nasalization in vowels (Pruthi, 2007) will be integrated along with an improved formant tracker Xia and Espy-Wilson (2000). Studies of pronunciation variability derived from previous work (Zhang, 1998) as well as continuing research will be integrated into EBS.

TABLE XI. Percent of TIDIGITS isolated digits with fully accurate broad class sequence.

| EBS (RBF) | HMM-AP | HMM-MFCC |
|---|---|---|
| 68.7 | 67.6 | 63.8 |

## ACKNOWLEDGMENTS

This work was supported by Honda Initiation Grant No. 2003 and NSF Grant No. BCS-0236707. The authors would like to thank Om Deshmukh at the University of Maryland for help with the HMM experiments.

[1] $F3$ was computed as an average over the third formant values obtained in voiced regions using on the ESPS formant tracker (Entropic, 1997). The same average value of $F3$ was used in each speech frame for computation of manner APs.

Ali, A. M. A. (**1999**). "Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition," Ph.D. thesis, University of Pennsylvania.

Bitar, N. (**1997**). "Acoustic analysis and modeling of speech based on phonetic features," Ph.D. thesis, Boston University.

Bitar, N., and Espy-Wilson, C. (**1996**). "A knowledge-based signal representation for speech recognition," *International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, 29–32.

Burges, C. (**1998**). "A tutorial on support vector machines for pattern recognition," Data Min. Knowl. Discov., **2**, 2, 121–167.

Carbonell, N., Fohr, D., and Haton, J. P. (**1987**). "Aphodex, an acoustic-phonetic decoding expert system," Int. J. Pattern Recognit. Artif. Intell. **1**, 31–46.

Chang, S. (**2002**). "A syllable, articulatory-feature, stress-accent model of speech recognition," Ph.D. thesis, University of California, Berkeley.

Chomsky, N., and Halle, N. (**1968**). *The Sound Pattern of English* (Harper & Row, New York).

Chun, R. (**1996**). "A hierarchical feature representation for phonetic classification," Master's thesis, Massachusetts Institute of Technology.

Clarkson, P., and Moreno, P. J. (**1999**). "On the use of support vector machines for phonetic classification," *International Conference on Acoustics, Speech and Signal Processing*, Phoenix, AZ, 485–488.

Clements, G. N. (**1985**). "The geometry of phonological features," Phonology Yearbook **2**.

Delgutte, B., and Kiang, N. Y. S. (**1984**). "Speech coding in the auditory nerve: Iv. sounds with consonant-like dynamic characteristics," J. Acoust. Soc. Am. **75**, 897–907.

Deng, L., and Sun, D. X. (**1994**). "A statistical framework for automatic speech recognition using the atomic units constructed from overlapping articulatory features," J. Acoust. Soc. Am. **100**, 2500–2513.

Deshmukh, O., Espy-Wilson, C, and Juneja, A. (**2002**). "Acoustic-phonetic speech parameters for speaker independent speech recognition," *International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, 593–596.

Deshmukh, O., Espy-Wilson, C, and Salomon, A. (**2005**). "Use of temporal information: Detection of the periodicity and aperiodicity profile of speech," IEEE Trans. Speech Audio Process. **13**, 776–786.

Eide, E., Rohlicek, J., Gish, H., and Mitter, S. (**1993**). "A linguistic feature representation of the speech waveform," *International Conference on Acoustics, Speech and Signal Processing* **93**, Minneapolis, MN, 483–486.

Entropic (**1997**). "Entropic signal processing system 5.3.1," Company out of business.

Espy-Wilson, C. (**1987**). "An acoustic phonetic approach to speech recognition: Application to the semivowels," Ph.D. thesis, Massachusetts Institute of Technology.

Glass, J. (**1984**). "Nasal consonants and nasalized vowels: An acoustic study and recognition experiment," Master's thesis, Massachusetts Institute of Technology.

Glass, J., Chang, J., and McCandless, M. (**1996**). "A probabilistic framework for feature-based speech recognition," *International Conference on Spoken Language Processing*, Philadelphia, PA, 2277–2280.

Halberstadt, A. K. (**1998**). "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. thesis, Massachusetts Institute of Technology.

Hasegawa-Johnson, M. (**1996**). "Formant and burst spectral measurements with quantitative error models for speech sound classification," Ph.D. thesis, Massachusetts Institute of Technology.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., and Wang, T. (**2005**). "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," *IEEE International Conference on Acoustic, Speech and Signal Processing*, Philadelphia, PA, 213–216.

Hirrsh, H. G., and Pearce, D. (**2000**). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR2000*, Paris, France, 181–188.

Hosom, J. P. (**2000**). "Automatic time alignment of phonemes using acoustic-phonetic information," Ph.D. thesis, Oregon Graduate Institute of Science and Technology.

Howitt, A. W. (**2000**). "Automatic syllable detection for vowel landmarks," Ph.D. thesis, Massachusetts Institute of Technology.

Joachims, T. (**1998**). "Making large-scale support vector machine learning practical," Advances in Kernel Methods: Support Vector Machines.

Juneja, A. (**2004**). "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. thesis, University of Maryland, College Park.

Juneja, A., and Espy-Wilson, C. (**2002**). "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," *International Conference on Neural Information Processing*, Singapore.

Juneja, A., and Espy-Wilson, C. (**2003**). "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," *International Joint Conference on Neural Networks*, Portland, OR, 675–679.

Juneja, A., and Espy-Wilson, C. (**2004**). "Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition," *From Sound to sense:* 50+ *Years of Discoveries in Speech Communication* (MIT, Cambridge MA), pp. C–151 to C–156.

Jurafsky, D., and Martin, J. H. (**2000**). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Prentice–Hall, Englewood Cliffs, NJ).

Keshet, J., Chazan, D., and Bobrovsky, B. (**2001**). "Plosive spotting with margin classifiers," Proceeding of Eurospeech, **3**, 1637–1640.

Kirchhoff, K. (**1999**). "Robust speech recognition using articulatory information," Ph.D. thesis, University of Bielefeld, Germany.

LDC (**1982**). "A speaker-independent connected-digit database," Http://www.ldc.upenn.edu/Catalog/docs/LDC93S10/, last viewed on January 30, 2007.

Lee, S. (**1998**). "Probabilistic segmentation for segment-based speech recognition," Master's thesis, Massachusetts Institute of Technology.

Liu, S. A. (**1996**). "Landmark detection for distinctive feature based speech recognition," J. Acoust. Soc. Am. **100**, 3417–3430.

NIST (**1990**). "Timit acoustic -phonetic continuous speech corpus," NTIS Order No. PB91 -5050651996.

Niyogi, P. (**1998**). "Distinctive feature detection using support vector machines," *International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, 425–428.

Pruthi, T. (**2007**). "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," Ph.D. thesis, University of Maryland, College Park.

Pruthi, T., and Espy-Wilson, C. (**2003**). "Automatic classification of nasals and semivowels," *International Conference on Phonetic Sciences*, Barcelona, Spain.

Rabiner, L., and Juang, B. (**1993**). *Fundamentals of Speech Recognition* (Prentice–Hall, Englewood Cliffs, NJ).

Salomon, A. (**2000**). "Speech event detection using strictly temporal information," Master's thesis, Boston University.

Salomon, A., Espy-Wilson, C, and Deshmukh, O. (**2004**). "Detection of speech landmarks from temporal information," J. Acoust. Soc. Am. **115**, 1296–1305.

Shimodaira, H., Noma, K., Nakai, M., and Sagayama, S. (**2001**). "Support vector machine with dynamic time-alignment kernel for speech recognition," Proceeding of Eurospeech, **3**, 1841–1844.

Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.

Vapnik, V. (**1995**). *The Nature of Statistical Learning Theory* (Springer-Verlag, Berlin).

Xia, K., and Espy-Wilson, C. (**2000**). "A new formant tracking algorithm based on dynamic programming," *International Conference on Spoken Language Processing*, Beijing, China, **3**, 55–58.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2006**). *HTK Documentation* (Microsoft Corporation and Cambridge University Engineering Department), http://htk.eng.cam.ac.uk/, last viewed January 30, 2007.

Zhang, Y. (**1998**). "Towards implementation of a feature-based lexical-access system," Master's thesis, Massachusetts Institute of Technology.

Zue, V., Glass, J., Philips, M., and Seneff, S. (**1989**). "The MIT summit speech recognition system: A progress report," DARPA Speech and Natural Language Workshop, pp. 179–189.