Detection of speech landmarks: Use of temporal information

Ariel Salomon,[a] Carol Y. Espy-Wilson, and Om Deshmukh
Electrical and Computer Engineering Department, University of Maryland, A.V. Williams Building, College Park, Maryland 20742

(Received 26 July 2001; accepted for publication 11 December 2003)

Studies by Shannon *et al.* [Science, **270**, 303–304 (1995)], Van Tasell *et al.* [J. Acoust. Soc. Am. **82**, 1152–1161 (1987)], and others show that human listeners can understand important aspects of the speech signal when spectral shape has been significantly degraded. These experiments suggest that temporal information is particularly important in human speech perception when the speech signal is heavily degraded. In this study, a system is developed that extracts linguistically relevant temporal information that can be used in the front end of an automatic speech recognition system. The parameters targeted include energy onset and offsets (computed using an adaptive algorithm) and measures of periodic and aperiodic content; together these are used to find abrupt acoustic events which signify landmarks. Overall detection rates for strongly robust events, robust events, and weak events in a portion of the TIMIT test database are 98.9%, 94.7%, and 52.1%, respectively. Error rates increase by less than 5% when the speech signals are spectrally impoverished. Use of the four temporal parameters as the front end of a hidden Markov model (HMM)-based system for the automatic recognition of the manner classes ''sonorant,'' ''fricative,'' ''stop,'' and ''silence'' results in the same recognition accuracy achieved when the standard 39 cepstral-based parameters are used, 70.1%. The combination of the temporal parameters and cepstral parameters results in an accuracy of 74.8%. [DOI: 10.1121/1.1646400]

PACS numbers: 43.72.Ar, 43.72.Ne [DOS] Pages: 1296–1305

I. INTRODUCTION

This paper investigates the use of temporal information for extraction of linguistically relevant details from a speech signal. This study has been considered in the context of a longer-term program on lexical access from features (LAFF) (Stevens, 1986; 2002) and event-based speech recognition (EBS) (Espy-Wilson, 1994; Juneja and Espy-Wilson, 2003), which is a furthering of the lexical access from spectra (LAFS) proposed by Klatt (1979). LAFF and EBS are paradigms for (human or machine) speech recognition in which landmarks in a speech signal are first located, and then features are attached to those landmarks. In this paper, motivated by studies of speech perception under certain types of degradation, we have concentrated on access to temporal information. The goal of this work is to develop a system that automatically extracts temporal cues that can be used in the front end of an automated speech recognition system.

A. Background

The motivations for this work come most importantly from studies of speech perception. A major problem in the development of speech recognition systems is the detection of speech from noise (cf. Viikki, 2001) or otherwise reduced spectral information. Recent studies show that human listeners can understand some aspects of the speech signal even when spectral shape has been significantly degraded. A source of information that may be of use, particularly under heavily degraded conditions, is that of temporal cues—information derived from the temporal structure of the speech signal. These cues are not targeted by traditional speech recognition systems, which generally focus on spectral features using data-derived spectral templates. Temporal processing is a significant factor in the auditory system, as observed by effects such as the phase locking of auditory-nerve firing to periodic signals (Moore, 1997). As such, this information should be available for use in human speech recognition.

The ability of human listeners to understand spectrally degraded speech has been examined in several studies (Shannon *et al.*, 1995; Van Tasell *et al.*, 1987; Turner *et al.*, 1995, among others) which demonstrate the ability of human listeners to recognize speech—particularly manner, nasality, and voicing—from primarily temporal cues. The general result has been that spectrally degraded speech still contains limited information about manner, nasality, and voicing. In light of these results, it is proposed that it should be possible to build a detector for acoustic events (prototypical landmarks) that is both largely independent of detailed spectral information and resistant to noise. Further, addition of temporal parameters should also improve performance and increase noise resistance for a system based on spectral information. Note, however, that improved performance may not occur if the noise has speech-like temporal characteristics. Turner *et al.* (1995) have shown that the use of temporal information is prevented when the speech signal is corrupted by modulated babble noise.

B. Temporal information

For the purpose of this study, it is helpful to specify exactly what is considered to be the ''temporal information''

[a]Currently at MIT, Research Lab of Electronics, Speech Communication Group, 77 Massachusetts Avenue, Rm. 36-511, Cambridge, MA 02142. Electronic mail: ariel@speech.mit.edu

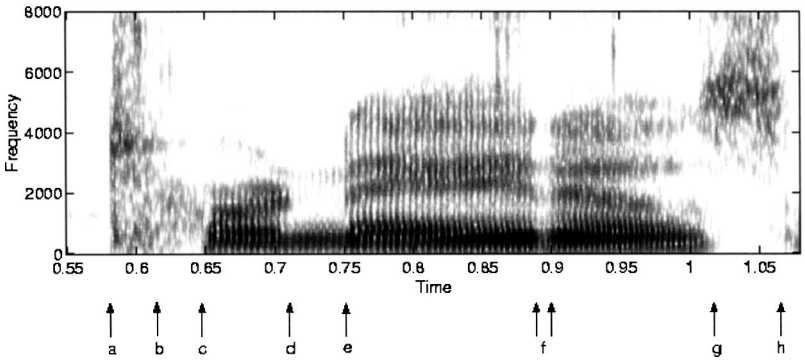

FIG. 1. Sample landmarks in the word ''tornadoes.'' (a) Release of stop consonant; (b) frication/aspiration boundary (nonrequired event); (c) onset of vowel; (d) closure for nasal consonant; (e) release of nasal consonant; (f) closure and release of stop consonant produced as a flap; (g) closure for fricative consonant; (h) release of fricative consonant.

in the signal. Temporal information is defined here in terms of bandpass components of the speech signal. Rosen (1992) proposes three categories of temporal information in speech: (1) ''envelope information'' (with fluctuations at rates from 2 to 50 Hz) which contains amplitude and duration cues to manner of articulation and voicing, as well as information about vowel identity (for example vowel length) and prosodic cues; (2) ''periodicity information'' (fluctuations at rates from approximately 50 to 500 Hz) which provides cues to voicing which can aid in manner identification, as well as marking stress locations by changes in pitch; and (3) ''fine structure'' (fluctuations at higher rates) which largely provides information also available from spectral shape. Note that normal-hearing subjects cannot detect amplitude fluctuations above about 1000 Hz; and response to modulation degrades rapidly above 100 Hz (Viemeister, 1979). This suggests that, at most, human listeners can only derive first-formant information from the temporal fine structure, if any information regarding fluctuations above the rate of pitch modulations is perceptually significant. Only the first two categories of temporal information are considered in this study.

Our goal in using temporal information is to aid in the analysis of the speech signal. In the human speech production system, a sequence of discrete elements (phonemes, words, etc.) is translated into an analog, continuous acoustic signal by the vocal apparatus. The process of understanding speech can be considered to be a reconstruction of the discrete stream of symbols from the speech signal. However, not all regions of a signal have the same information content: steady-state portions can be sampled slowly to determine overall properties, while abrupt points such as the critical point of a transition for a consonantal closure can contain a significant number of linguistically important cues in a concentrated region. These transition regions can contain information about the adjacent segments, most importantly in the type of transitions in and out of the target position. In this paper, we are concerned with landmarks involving abrupt acoustical changes, a set of which is illustrated for an utterance in Fig. 1.

C. System goals

The goal of event detection is to generate a set of landmarks referred to as *events* that will direct further analysis of the speech signal. To ensure the success of further levels of processing (outside the scope of this paper), this set should be essentially complete with respect to the perceptually sharpest events, for example events corresponding to stop-consonant bursts, strident fricatives, and stressed vowels. Note that insertions are somewhat less critical as they can be discarded by further analysis. On the other hand, it is likely that some weaker events are going to be captured less often: semivowels (particularly the glides /w/ and /y/), for which the primary cues consist of formant movement (Espy-Wilson, 1992); weak fricatives which have become sonorant, such as a common pronunciation of the /v/ in ''everyday'' (Catford, 1977; Espy-Wilson, 1994; Deshmukh and Espy-Wilson, 2003) and other cases of events that do not involve a significant degree of energy fluctuation. In cases of heavily coarticulated segments, it is expected that the output of the system will reflect the type of events that actually occurred rather than the canonical events expected from segment-based labels (e.g., sonorant events rather than onset and offset of frication for the /v/ in ''everyday'' when it is manifest as a sonorant consonant).

The parameters used to locate landmarks in the speech signal are changes in spectral energy or in the periodicity content of the signal, corresponding to the two relevant types of temporal information discussed above. This work relies on the assumption that abrupt landmarks are initially located based on only amplitude changes in the signal (Stevens, 2002). It is at later stages of processing that spectral information is integrated, and then higher-level information such as phonotactics and lexical constraints is applied. The question in this study is how much information can be detected regarding manner information in speech from strictly temporal information. It is certain that additional use of spectral information, wider context, and high-level constraints will be able to improve the results.

This work builds on work by Espy-Wilson (1992), where the degree of abruptness in an energy difference measure was used to distinguish the glides /w,y,r/ from the consonantal sonorant consonants /m,n,l/, and by Bitar (1997), where energy differences were used to find stop bursts and to perform segmentation. The methodology is similar to research by Liu (1994), who used energy differences computed from specific frequency bands to find certain types of landmarks in the speech signal. In addition, it bears similarity to the processing used by Browne and Cooke (1994), who used onsets and offsets and pitch information as cues in a system for auditory scene analysis. In this study, event detection is based on general processing of all frequency bands and the

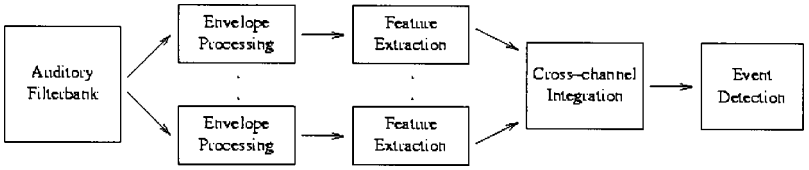

FIG. 2. Overall structure of analysis algorithm. Several stages of computation are performed within each channel, followed by integration into summary measures for use in event detection.

signal processing is adaptive. Energy change is combined with information about periodicity and aperiodicity to detect events; and analysis is performed for both clean and spectrally impoverished speech signals. In addition, we compare the performance of these temporal parameters with the traditional cepstral-based parameters in the manner classification of speech using an HMM-based recognition system (Young, 1995).

II. METHOD FOR LANDMARK DETECTION EXPERIMENTS

A. Database

The TIMIT database (Seneff and Zue, 1988) was used as a corpus of labeled speech data. This is a widely available database of speech recorded in quiet and labeled at the level of words and phonetic strings. Although it would have been more useful to use a database labeled at the landmark level (e.g., a database currently under development at the Massachusetts Institute of Technology; see Choi *et al.*, 1997), a large enough database of this type was not yet available. The TIMIT database consists of 6300 utterances spoken by 630 speakers, of which 4620 utterances make up the suggested training set and 1680 are in the test set. In particular, we used the phonetically compact (**sx**) sentences. Training was performed using a set of 20 **sx** sentences (spoken by 10 males, 10 females) randomly drawn from the TIMIT training set. Testing was performed using all 120 **sx** sentences from the TIMIT core test set (spoken by 8 female speakers and 16 male speakers, evenly distributed over all 8 dialect regions).

For the purpose of comparing with the detected landmarks, a set of expected (posited) landmarks was generated from the phonetically labeled transcriptions available in the TIMIT database using a simple rule-based algorithm based on the manner class of adjacent segments at each boundary. The posited landmarks are expected to have some inherent error as the mapping is underspecified. Some of the underspecification in the TIMIT labels is accounted for by inserting events that are labeled as ''nonrequired'' because they are possible, and may be caught by the matching algorithm, but not necessarily strongly expected. An example of a nonrequired event is the frication/aspiration boundary (event b) in Fig. 1. [Note from part (e) of Fig. 4 that this boundary is detected as a −C event by our algorithm.] The 20 utterances in the training database were also hand labeled for more reliable system development and training. For analysis of the effect of using generated landmark labels, the trained system was evaluated on both the generated labels and the hand labels of the training set. The overall error rate was 18.5% for the training set with the generated labels and 14.8% for the training set with the hand labels. Focusing on only the robust set of landmarks, the error rate was 6.54% for the training set with the generated labels and 6.06% for the training set with the hand labels. Note that the robust set contained approximately 70% of the total set of landmarks under analysis.

In addition to using this clean version of TIMIT to test our algorithm, we also spectrally impoverished the TIMIT database to see how well the temporal parameters perform with degraded spectral information. The spectral impoverishment was performed using a bank of four filters in a technique developed by Shannon *et al.* (1995).

B. Signal analysis

Signal analysis consisted of a series of stages per channel, as shown in Fig. 2. The signal was first filtered into a set of bandpass frequency channels, and each narrow-band signal was examined independently. In each channel, the signal underwent envelope analysis and feature extraction, which consisted of periodicity measurements and an energy onset/offset measure. This was followed by combination into a number of cross-channel summary measures as functions of time: summary levels of aperiodicity and periodicity, pitch, and energy onset and offset measures. The resulting waveforms were analyzed to locate events in the speech signal.

The filter bank used was a 60-channel auditory gammatone filter bank with characteristic frequencies (CFs) based on the ERB scale (Patterson, 1992). An auditory filter bank was chosen for spectral analysis in order to provide an accurate weighting of frequency components, most importantly in terms of the strength of events corresponding to voiced excitation of speech relative to their unvoiced counterparts.

In order to avoid excessive smoothing in the time domain, an envelope operator based on the Hilbert information (Rabiner and Gold, 1975) was used. The envelopes $e_i(t)$ of the individual channels are obtained by the function

$$e_i(t) = |x_i(t) + j \cdot H\{x_i(t)\}|,$$

where $x_i(t)$ is the input signal, and $H\{x_i(t)\}$ is the Hilbert transform of the input signal. Given a real narrow-band signal as input, the Hilbert transform produces a version of its input signal that is precisely 90° out of phase, such that the amplitude of the complex sum of these two signals is an estimate of the low-frequency amplitude modulation applied to the signal. This transform is an improvement over a simple smoothing operation because abrupt changes are preserved, at the maximum rate that can be captured by a particular channel given its CF.

1. Periodicity and aperiodicity feature extraction

The first feature extraction algorithm applied to the temporal envelopes makes a three-way classification between silence, periodic, and aperiodic in each channel every 2.5 ms.

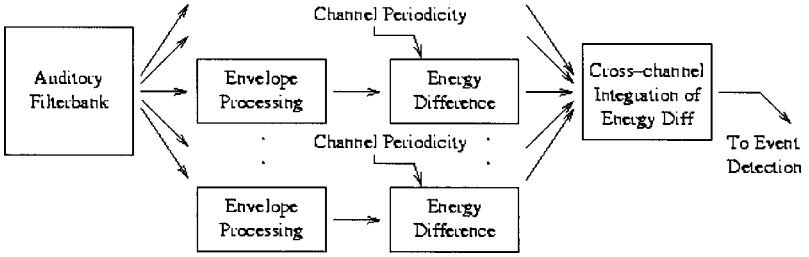

FIG. 3. Structure of energy analysis component of feature extraction. Note that the first difference operators used as the energy difference measure are adapted based on periodicity information within each channel. Following this, summary onset and offset measures are computed across all channels.

A periodic signal is defined to be one that contains regular temporal fluctuations at rates from roughly 55–500 Hz (Deshmukh and Espy-Wilson, 2003).

2. Energy operator feature extraction

The other major feature extraction algorithm used is an energy onset/offset detector based on a first-difference measure, originally derived from the onset/offset parameters designed by Espy-Wilson (1992) to capture rate of spectral change for distinguishing between sonorant consonants. These parameters were later used by Bitar (1997) to detect the abrupt onset of stop consonants. The onset/offset measure in this study is constructed from first differences in each channel output from the temporal envelope analysis stage described above [see parts (d) and (e) of Fig. 4]. The first difference is computed as a log difference between the sum amplitude of two adjacent nonoverlapping windows of the signal in a particular channel, as per the formula

$$D_{i,k} = 20 \log \sum_{m=-\infty}^{\infty} x_i(n+m)w(m) - 20 \log \sum_{m=-\infty}^{\infty} x_i(n+m-k)w(m-k),$$

where $x_i(n)$ is an individual channel input signal, k is the time difference between the two windows, and the windows $w(n)$ are rectangular windows of length k. The computed difference is scaled in decibels (dB). This first difference operation is essentially the same as the rate-of-rise (ROR) detector used by Liu (1994), but the two windows are adjacent in time to minimize error in location of the detected peaks.

It was observed that by increasing the window sizes (and correspondingly increasing k, referred to as the difference time) for the first difference computation, noise in the measurement over the utterance is reduced, particularly in fricative regions. However, an unfortunate side effect of lengthening the window sizes was a decrease in the strength of peaks and temporal accuracy in the onset signal associated with stop bursts. In order to obtain the advantages of a long window size, a method of dynamically adapting the difference time based on features of the signal measured by the periodicity detector was developed. Under this method, the energy difference detector is adapted in each channel independently, with difference length targets based on the existence of silence or periodic/aperiodic excitation, and according to the pitch estimate in periodic regions, as follows: (1) the difference time is shortened (5 ms) for silence to sharpen the response to onsets that are preceded by silence (as expected for stop bursts); (2) the difference time is lengthened (30 ms) in aperiodic regions, to maximally smooth the first difference output in fricated segments; and (3) the difference time is tuned to exactly twice the pitch period in periodic regions, to prevent detection of spurious energy fluctuation due to misalignment with the pitch period. There is also a slew rate control of 0.5 ms per millisecond (the difference operator is sampled every ms) to prevent discontinuities.

3. Summary measures

The measurements made in individual channels are combined to produce summary measures. The silence/periodic/aperiodic decisions are combined across channels to produce two measurements called P_{eng} and AP_{eng}, which are the proportions of periodic energy and aperiodic energy in the signal, respectively.

From the per-channel differences, two measures are computed: the positive differences (increasing levels) are summed to produce an ''onset'' signal, and the negative differences (decreasing levels) are summed to produce an ''offset'' signal. The offset parameter is usually inverted for analysis to make it positive, allowing generalization of all further computations; note, however, that the noninverted negative version of the parameter is the one shown in all figures. A scaling by $1/N$, where N is the total number of channels, produces an average energy change per channel on a dB scale

$$\text{on}(n) = \frac{1}{N} \sum_{i:D_{i,k}(n)>0} D_{i,k}(n),$$

$$\text{off}(n) = \frac{1}{N} \sum_{i:D_{i,k}(n)<0} D_{i,k}(n).$$

This set of parameters in combination over a speech signal visibly provides useful information about the content of the signal, as can be seen in Fig. 4 for the same utterance used in Fig. 1. Note that the periodicity and aperiodicity proportion measures in (b) provide a decomposition of the signal into periodic (roughly, voiced) and aperiodic elements. Also note that the onset and offset measures in (d) have peaks at most of the important events in the signal.

C. Event detection

The manner classes of speech segments are listed in Table I, along with corresponding source types. Derived from these classes for the purpose of detection, a set of event types based on acoustic parameters was defined, and is listed in Table II. The categories correspond to the polarity (onset or offset of energy) of the event, and their correlation with periodic and/or aperiodic excitation. Events are labeled

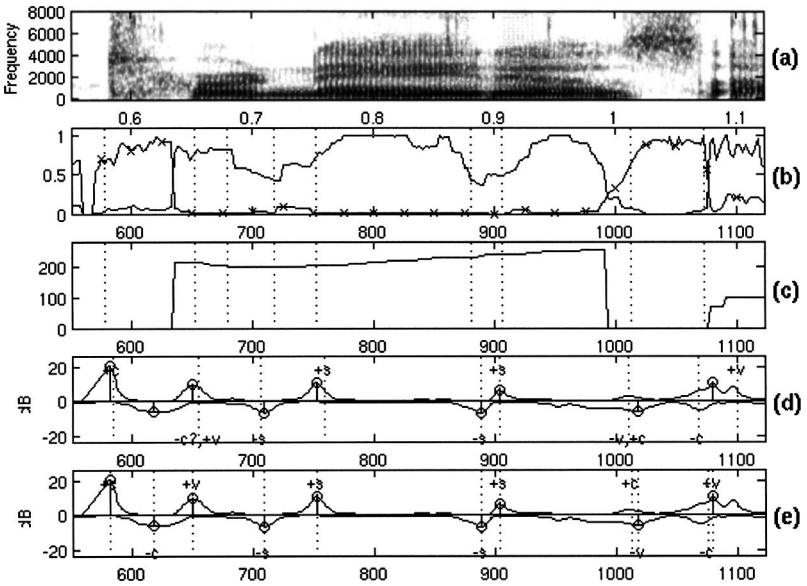

FIG. 4. Parameter extraction results for the word "tornadoes." (a) Spectrogram; (b) proportion of periodic (solid line) and aperiodic (solid line with "×" overlaid) energy; (c) smoothed pitch estimate in periodic regions; (d) onset and offset parameters, chosen peaks (stems), posited events (nonrequired marked with ?); (e) detected events.

based on their occurrence either at a boundary where periodic content begins or ends (±V, correlated with voicing onset or offset), or when surrounded by periodic excitation (±S, correlated with sonorant consonant boundaries), or their occurrence at a boundary of aperiodic excitation or at least occurring outside of periodic excitation (±C, correlated with obstruent consonants). The output of the event detector consists of this set of event labels. Part (d) of Fig. 4 shows a posited set of these events generated from TIMIT labels; part (e) shows detected events from the speech signal.

The normalized summary periodic and aperiodic energy measures, P_{eng} and AP_{eng}, are analyzed (after median smoothing) to locate potential confident regions and their boundaries. The set of confidently periodic/aperiodic regions is determined by applying a minimum threshold for a required maximum value of P_{eng} or AP_{eng} for a region. Following this, lower thresholds are used to find the boundaries of each region. Aperiodic regions are discarded unless at least one end of the region is associated with an onset/offset event, i.e., the beginning of the region near an onset event or the end of the region near an offset event. They are also discarded if the aperiodic region is shorter than 10 ms.

The onset and offset parameters are converted into a sequence of potential events by use of a convex hull-based peak-picking algorithm. There are thresholds for minimum peak height, and a required minimum dip between two adjacent peaks. Onset and offset peaks are associated with boundaries of periodic/aperiodic regions in order to classify event types. Onset/offset peaks located near the beginning/end of a periodic region are labeled as ±V. Correspondingly, onset/offset peaks located near the beginning/end of an aperiodic region are labeled as ±C. The locality criteria are determined from trained thresholds. Remaining boundaries of confidently periodic/aperiodic regions are labeled as landmarks of the corresponding types, but note that the times are less accurate.[1] Remaining onset/offset peaks are labeled as ±S if they are within a periodic region, or ±C if they are outside of any periodic region. The full set of trained parameters used in this process is listed in detail in Table III.

TABLE I. Modes of speech production (manner classes).

Manner	Oral tract	Primary source
Vowel	Open	Periodic
Semivowel	Slightly constricted	Periodic
Nasal	Closed (with nasal coupling)	Periodic
Fricative	Narrow constriction	Aperiodic
Stop	Completely closed	Aperiodic

D. Training procedure

Some adjustment was performed on a number of the time, energy, and confidence-level thresholds involved in event extraction. These included the pairs of thresholds used for determining confident regions of periodicity or aperiodicity as discussed above. The training procedure adjusted a set of 12 parameters, listed in Table III. The procedure was a sequence of Nelder–Mead simplex optimization stages on three subsets of the parameters (as defined in Table III), performed over the score $S = N_{\text{matches}} - N_{\text{insertions}}$ (Nelder and Mead, 1965). This cost is equivalent to minimizing the total error rate, as the base number of posited required events will not change (and as such a decrease in the number of matches corresponds to an increase in the number of deletions). The

TABLE II. Event types.

Label	Name	Description
+V	Voicing onset	Onset corresponding to beginning of periodicity (beginning of a vowel or sonorant consonant)
−V	Voicing offset	Offset corresponding to end of periodicity (end of a vowel or sonorant consonant)
+S	Sonorant onset	Onset within periodic region (onset at release of nasal or semivowel)
−S	Sonorant offset	Offset within periodic region (offset at release of nasal or semivowel
+C	Obstruent onset	Onset corresponding to beginning of aperiodicity (stop consonant burst, affricate or fricative onset)
−C	Obstruent offset	Offset corresponding to end of aperiodicity (stop, affricate or fricative offset)

TABLE III. Parameters with trained values. P_{on} and P_{off} refer to the boundaries of a periodic region; AP_{on} and AP_{off} are the corresponding locations for an aperiodic region.

Parameter	Description	Value
	Periodicity parameters	
$t_{before}:P_{on}$	Max. time from P_{on} to corresp. onset peak (peak *precedes* P_{on})	59.8 ms
$t_{after}:P_{on}$	Max. time from P_{on} to corresp. onset peak (peak *follows* P_{on})	4.48 ms
T_{per_RGN}	''Peak threshold'' on P_{eng} to consider a region as periodic	58.7%
T_{per}	''Boundary threshold'' on P_{eng} to located ends of a periodic region	31.1%
$t:P_{off}$	Maximum time between P_{off} and corresponding offset peak	61.7 ms
	Aperiodicity parameters	
$t:AP$	Max. time between AP_{on}/AP_{off} and corresponding on/off peak	31.1 ms
T_{aper_RGN}	''Peak threshold'' on AP_{eng} to consider a region as aperiodic	84.2%
T_{aper}	''Boundary thresh.'' on AP_{eng} to located ends of aperiodic region	66.0%
	Onset/offset parameters	
T_{on_peak}	Minimum peak height in onset measure	4.70
T_{on_dip}	Minimum dip between peaks in onset measure	4.70
T_{off_peak}	Minimum peak height in offset measure	5.15
T_{off_dip}	Minimum dip between peaks in offset measure	5.15

procedure was dependent on initial conditions, which were set by trial and error and knowledge of front-end behavior. The training process was iterated twice to ensure some degree of convergence.

E. Scoring algorithm

A standard algorithm used for scoring speech recognizer performance at the phonetic level was modified to support scoring landmark results. The algorithm was derived from the DARPA speech recognizer performance evaluation tools (Pallett, 1989). This code aligns a recognized token string with reference labels using a dynamic programming algorithm. The original code supported scoring costs for insertions, deletions, and substitutions in a stream of labels. Modifications were made to perform the task of landmark scoring: (a) a cost was added for the difference in time (in ms) from the posited label to the detected label, to ensure that label matches and substitutions were close in time (insertion/deletion costs are equivalent to the cost of matching a label 50 ms from its posited location); (b) support for nonrequired events with zero deletion cost was added; (c) support for pairs of co-occurring events which could be found in either order was added, for example the onset of a fricative at the same point as the offset of the preceding vowel; and (d) substitution cost was doubled in the case that the polarity was incorrect, such that +C for −C was a more costly substitution than −V for −C, as it was more likely in the polarity mismatch cases that there was actually both an insertion and a deletion, rather than just a substitution.[2] Additional adjustments in the final score were made to ignore insertions before the beginning and after the end of the labeled speech, under the assumption that integrating an endpoint detector in the system would prevent positing events at these locations.

A set of summary statistics was defined to analyze matching results. All are defined in terms of the base rate N, the number of posited tokens not counting neutral deletions (of tokens marked as non-required). Defining N_P as the total number of posited tokens (including those marked nonrequired), D as the number of error deletions (of required tokens), D_N as the number of neutral deletions, S as the number of substitutions, and I as the number of insertions, the metrics are computed according to the following formulas:

$$N = N_P - D_N \quad \text{(base rate of matched tokens)},$$

$$\text{detection rate } R_M = \frac{N - D - S}{N},$$

$$\text{deletion rate } R_D = \frac{D}{N},$$

$$\text{substitution rate: } R_S = \frac{S}{N},$$

$$\text{insertion rate: } R_I = \frac{I}{N}.$$

III. METHOD FOR BROAD-CLASS RECOGNITION EXPERIMENTS

A. Database

The TIMIT database was used for the recognition experiments. The training data consisted of 2710 utterances from the suggested training section of the TIMIT database. The performance of the recognizers was based on 504 utterances from the suggested test set.

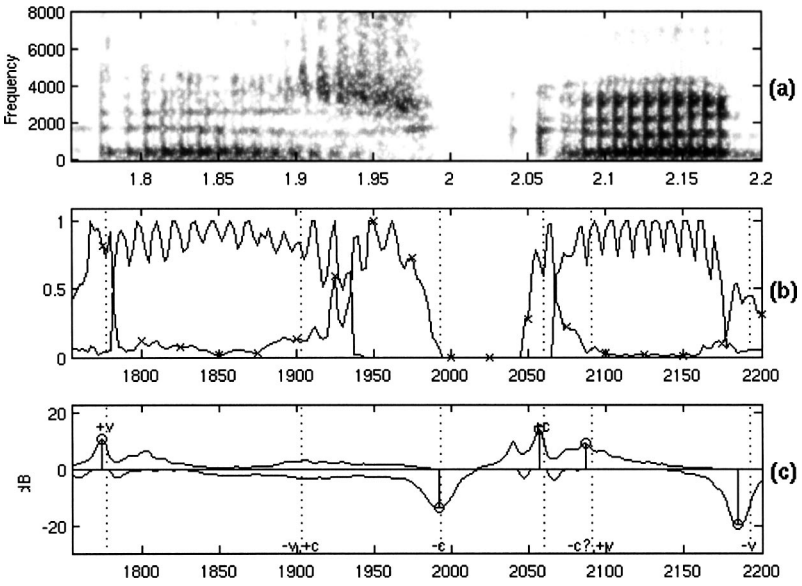

FIG. 5. Parameter extraction results for the fragment ''is going,'' displaying the difficulty of locating the boundary between a vowel and an adjacent voiced fricative /z/. (a) Spectrogram; (b) proportion of periodic (solid line) and aperiodic (solid line with ''×'' overlaid) energy; (c) onset and offset parameters, chosen peaks (stems), posited events.

B. Models

Acoustic models were built for four manner classes: sonorant (includes vowels and sonorant consonants), stop, fricative, and affricate. In addition, a model was built for silence. The back-end processing was done using an HMM-based recognition system (Young, 1995). Each model consisted of a three-state (plus 2 terminal states) HMM with eight Gaussian mixtures in each state. Each mixture was initialized as zero mean and unit diagonal covariance. Each mixture had diagonal variance and all the mixtures weights in all the states were initialized at the same value. Left-to-right state transition with one skip was incorporated with the additional constraint that each model had to start at the first state. All of the allowable transitions were initialized as equiprobable.

Three different front ends were used in the recognition experiments. The first front end consisted of 12 mel-frequency cepstral coefficients (MFCCs) and energy with their delta and acceleration coefficients. The second front end consisted of the four temporal-based parameters: aperiodic energy measure, periodic energy measure, onset waveform, and offset waveform. The third front end consisted of both the cepstral-based parameters and the temporal-based parameters. All of the parameters were computed at a rate of 5 ms and the window size was 20 ms. The mean of each parameter was normalized to zero.

C. Scoring

For scoring, the phonetic transcriptions provided with the TIMIT database were mapped into the four manner classes and silence. Although separate models were built for affricates and stops, they were recognized as the same class. Flaps were allowed to score as either a stop or a sonorant consonant. Glottal stops were allowed to score as either a vowel or a stop.

IV. RESULTS

A. Detection of events in clean and spectrally impoverished speech

Results for the event detection in clean and spectrally impoverished speech are plotted separately for the categories ''strongly robust,'' ''robust,'' and ''weak'' event types in Fig. 5. Weak events were a set of events that were expected to be less abrupt, including releases and closures for nasals, voiced nonstrident fricatives, and voiced stops labeled as flaps. The rest of the events were considered robust, and a subset of these that were detected with an error rate (includes deletions and substitutions) less than 2% in clean speech was labeled as strongly robust. Details of the detection rates for different events are given in the Appendix. Events were detected[3] with an overall detection rate of 80.2% on the clean test data set with an insertion rate of 8.7%, and 76% on the impoverished test set with an insertion rate of 36.6%. Note that for each category, the difference between the detection results for the clean speech and the impoverished speech is within 5%. Thus, the temporal parameters are quite robust to spectral degradation.

Nearly half of the error rate is due to missed landmarks at the boundary between a nasal consonant and a vowel, an event type that was detected with only 45.6% accuracy. Another major error source was from landmarks for voiced nonstrident fricatives; initial onsets preceding the fricative were located only 48.1% of the time, and landmarks for a voiced weak fricative adjacent to a vowel were detected with 49.0% accuracy. A third difficult case involved locating landmarks for stop consonants labeled as flaps, of which only 42.6% were detected correctly. These three cases combined account for 69.5% of all errors made. Discounting these classes of landmarks, the detection rate was 91.8%; and the detection rate for a subset consisting of the most strongly robust event types was 98.5%.

Landmark types that were detected well included stop consonants and unvoiced fricatives: 90.9% of stop closures following a vowel were detected, and 96.0% of stop bursts

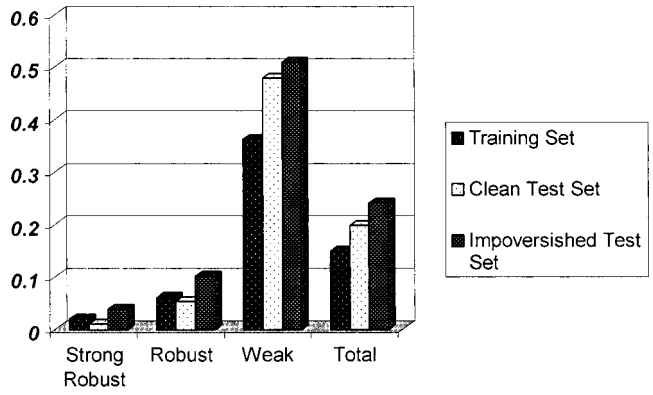

FIG. 6. Landmark detection results for clean and impoverished speech.

were detected (including 99.0% of unvoiced stop releases). Landmarks for unvoiced fricative closures and releases adjacent to a vowel were detected with 99.0% accuracy (and 92.4% for voiced stridents). Note that affricate consonants were grouped with strident fricatives for this count. The voiced /z/ and /zh/ fricatives can have a difficulty to locate boundary due to overlap with neighboring segments; see, for instance, the utterance shown in Fig. 6. More results for robust event types are given in the Appendix.

The results from this study are somewhat different in organization than those of Liu (1994), but a gross comparison of some of the results is possible. When tested across the full TIMIT test set using automatically generated labels, the landmark detector developed by Liu had an error rate of 15% for deletions, 6% for substitutions, and 25% for insertions. Our overall error rate on a subset of the TIMIT test data is 19% for deletions when laterals are included (15% when laterals are not included in the events expected to be detected, 4.8% for substitutions, and 8.7% for insertions). Note that these numbers may not be directly comparable since there are differences in the way the results were tallied. One possible conclusion from this comparison is that a selection of broader frequency bands such as those used by Liu may be more appropriate for the detection of nasals and laterals, whereas summary measures across all frequency channels may be better for obstruents.[4]

B. Temporal parameters vs cepstral parameters for manner-class recognition

The manner class recognition results are given in Table IV. As can be seen, the four temporal parameters result in performance that is comparable to that obtained with the 39 cepstral-based parameters. Adding the temporal-based parameters to the cepstral-based parameters results in close to a 5% improvement in accuracy. This increase in performance is due largely to improved detection of landmarks, particularly for the stop and fricative consonants. Note that the performance of the recognizers may in fact be considerably better than the numbers in Table IV suggest. In an analysis of manner–class recognition experiments performed by Bitar and Espy-Wilson (1995) using some of the same TIMIT test sentences, Bitar and Espy-Wilson found that presumed misclassifications of voiced obstruents as sonorant consonants was not incorrect. Many of the voiced obstruents were in fact realized as sonorant consonants, even though this change in the surface realizations of the voiced obstruents is not reflected in the TIMIT transcriptions. More recent results evaluating the proportion of periodic vs aperiodic energy in speech signals by Deshmukh and Espy-Wilson (2003) show that about 33% of voiced obstruents have a strong periodic component with little or no aperiodic component, and about half of all voiced obstruents show strong periodicity.

TABLE IV. Recognition results (in percent) for the broad classes: Sonorant (includes vowels and sonorant consonants), fricative, stop and silence.

	Correct	Accuracy
MFCCs (39 parameters)	73.9	70.1
Temporal measures (4 parameters)	78.0	70.1
MFCCs+temporal measures (43 parameters)	81.1	74.8

V. DISCUSSION

There are a number of key areas where accuracy could be improved, particularly in use of prediction and longer-term integration of information. In the front end, silence detection may be improved by addition of masking constraints; this will result in improved contextual reliability as a function of signal level. Of primary interest are spectral masking effects of tones on noise with respect to thresholds for detection of aperiodicity in the presence of a primarily periodic signal. The onset/offset detector could benefit from improved adaptation of the temporal sensitivity parameter; one possibility may be to examine separately adapting versions tuned to periodic excitation and aperiodic events for optimal detection of fricative events. Finally, it may be possible to modify the event extraction stage of the system to dynamically adapt thresholds as necessary. This could be done using temporal constraints (expected events or segments per unit time, adapted for speaking rate) rather than explicit required levels for peaks.

An important area for further research will be improving the extraction of temporal cues from noisy inputs. The present system is subject to errors (likely to be primarily insertions, which may be filterable based on spectral characteristics) given a signal mixed with rapidly fluctuating noise. The use of a large number of narrow bands allows for adaptation to noise if methods are developed to identify which bands to ignore; such methods could include correlation methods, or scene analysis techniques such as those used in missing-data approaches to speech recognition (Cooke *et al.*, 1997).

The next stage of this work will be to combine cues derived from temporal information with a recognition system based on spectral features, as both types of features will be important for a complete system. It is clear that temporal information is used by the human speech recognition system, and so should be critical to achieving high-quality performance in a computer speech recognition system; spectral cues are also of significant importance, for example formant frequencies and other spectral cues to place. This merging

could involve spectral weighting of temporal components, or a merger at the event output stage, increasing confidence in existence of an event if it is posited from multiple types of information. Later stages of the recognizer will be able to take into account a broader range of temporal and spectral cues.

VI. CONCLUSION

This work has shown that use of temporal information for landmark detection is feasible, particularly for a subset of robust abrupt events such as stop bursts. Although previous studies have investigated the use of temporal information in particular cases or as an additional measure, this work extends this body of work by using temporal information everywhere as the primary information source. As noted by use of a tunable onset/offset detector, it was determined that some locations require different degrees of sensitivity to temporal information. It has also pointed to certain landmark types where spectral features and perhaps more subtle temporal features (on a longer time scale) are important, particularly for landmarks related to sonorant consonants. It would be expected that the optimal system would integrate both temporal and spectral information.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant No. SBR-9729688 and NIH Grant No. 1K02 DC00149-01A1. Thanks to the three anonymous reviewers and the associate editor for useful comments on this manuscript.

APPENDIX A: ERROR RATES IN THE DETECTION OF ROBUST EVENTS IN THE CLEAN TEST SET

Landmark type	% Deletion and substitution errors	% Deletion error	# of tokens
	Consonantal landmarks adjacent to vowels		
Stop closure	9.14	1.57	383
Unvoiced stop release	1.03	0.34	291
Voiced stop release	8.51	0.00	188
Unvoiced fricative	1.01	1.01	397
Voiced fricative	7.05	7.05	156
	Interconsonantal landmarks		
Stop/strident fricative boundary	1.23	0.00	81
Strident fricative/stop boundary	5.49	4.40	91
Fricative/nasal and nasal/fricative boundary	5.00	5.00	40
	Initial and final events		
Glottal onset/offset	10.50	5.88	238
Strident fricative (initial or final)	1.11	1.11	90
Unvoiced weak fricative (initial or final)	11.11	11.11	9

APPENDIX B: ERROR DETECTION RATE OF WEAK EVENTS IN CLEAN TEST SET

Landmark type	% Substitution and deletion errors	% Deletion errors	# of tokens
	Consonantal landmarks adjacent to a vowel		
Stop (labeled as a flap)	57.45	31.91	94
Stop release (labeled without a burst)	46.67	13.33	15
Voiced weak fricative	51.03	51.03	194
Nasal	54.45	43.64	472
Aspiration (h)	31.82	31.82	44
	Interconsonantal landmarks		
Stop/nonstrident fricative or nonstrident fricative/stop boundary	19.05	14.29	42
Non-strident fricative/nasal or nasal/nonstrident fricative boundary	33.33	33.33	27
Stop/nasal or nasal/stop boundary	21.18	20.00	85
	Initial and final events		
Initial voiced nonstrident fricative	51.85	51.85	27
Aspiration (initial or final)	36.36	36.36	11

[1]The periodicity results are compiled only every 2.5 ms, whereas onset/offset parameters are computed with a 1-ms frame rate. Also, the energy change measurement is inherently more accurate in time as the periodicity computation is dependent on even longer time scales corresponding to pitch periods.

[2]Due to inclusion of a cost for the distance in time between the posited and generated events, this type of substitution would never be chosen by the scoring algorithm, as the cost structure makes it cheaper for the system to count it as an insertion plus a deletion.

[3]Events for flaps (labeled ''dx'') were considered correct whether they were detected as ±V or ±S.

[4]Note that Liu provides detailed data similar to those listed in the Appendix, but it is for the LAFF database which was hand labeled. Although a direct comparison is difficult given the differences in the databases and the differences in the generation of the reference labels, Liu's detector may perform better on nasal consonants than our detector (29% error on closures and 44% error on releases, vs 54.5% error), but it does not reach our level of performance for unvoiced stop releases (8% error vs 1.0% error).

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

Bitar, N., and Espy-Wilson, C. Y. (**1995**). ''A signal representation of speech based on phonetic features,'' Proceedings of the IEEE Dual-Use Technologies and Applications Conference, SUNY Inst. of Tech., Utica, Rome, pp. 310–315.

Browne, G. J., and Cooke, M. (**1994**). ''Computational auditory scene analysis,'' Comput. Speech Lang. **8**, 297–336.

Catford, J. C. (**1977**). *Fundamental Problems in Phonetics* (Indiana University Press, Bloomington, IN).

Choi, J. Y., Chuang, E., Gow, D., Kwong, K., Shattuck-Hufnagel, S., Stevens, K. N., and Zhang, Y. (**1997**). ''Labeling a speech database with landmarks and features,'' 134th Meeting of the Acoustical Society of America, S3163.

Cooke, M., Morris, A., and Green, P. (**1997**). ''Missing data techniques for robust speech recognition,'' Proceedings of the ICASSP, pp. 863–866.

Deshmukh, O., and Espy-Wilson, C. (**2003**). ''Detection of the periodicity and aperiodicity profile of speech,'' Proceedings of the ICASSP, pp. 448–451.

Espy-Wilson, C. Y. (**1992**). ''Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English,'' J. Acoust. Soc. Am. **92**, 736–757.

Espy-Wilson, C. Y. (**1994**). ''A feature-based semivowel recognition system,'' J. Acoust. Soc. Am. **96**, 65–72.

Glasberg. B. R. and Moore, B. C. J. (**1990**). ''Derivation of auditory filter shapes from notched-noise data,'' Hear. Res. **47**, pp. 103–138.

Juneja, A., and Espy-Wilson, C. Y. (**2003**). ''An Event-based Acoustic Phonetic Approach for Speech Segmentation and E-Set Recognition,'' Proceedings of the International Congress of Phonetic Sciences, pp. 1333–1336

Klatt, D. (**1979**). ''Speech perception: A model of acoustic-phonetic analysis and lexical access,'' J. Phonetics **7**, 279–312.

Liu, S. (**1994**). ''Landmark detection for distinctive feature based speech recognition,'' Ph.D. thesis, Massachusetts Institute of Techology, Cambridge, MA.

Moore, B. C. J. (**1997**). *An Introduction to the Psychology of Hearing*, 4th ed. (Academic, London).

Nelder, J. A., and Mead, R. (**1965**). ''A simplex method for function minimization,'' Comput. J. (UK) **7**, 308–313.

Pallett, D. S. (**1989**). ''Benchmark tests for DARPA resource management database performance evaluations,'' in Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing, pp. 536–539.

Rabiner, L. R., and Gold, B. (**1975**). *Theory and Application of Digital Signal Processing* (Prentice Hall, Englewood Cliffs, NJ).

Rosen, S. (**1992**). ''Temporal information in speech: Acoustic, auditory, and linguistic aspects,'' Philos. Trans. R. Soc. Lond. B Biol. Sci. **336**, 367–373.

Seneff, S., and Zue, V. (**1988**). ''Transcription and alignment of the TIMIT database,'' included with TIMIT database.

Shamma, S., and Hermansky, H. (**2000**). ''Speech recognition from temporal patterns,'' Proceedings of the ICSLP.

Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). ''Speech recognition with primarily temporal cues,'' Science **270**, 303–304.

Stevens, K. N. (**1986**). ''Models of phonetic recognition. II. A feature-based model of speech recognition,'' in Proceedings of the Montreal Satellite Symposium on Speech Recognition, Twelfth International Conference on Acoustics, edited by P. Mermelstein, pp. 67–68.

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

Turner, C. W., Souza, P. E., and Forget, L. N. (**1995**). ''Use of temporal envelope cues in speech recognition by normal and hearing impaired listeners,'' J. Acoust. Soc. Am. **97**, 2568–2576.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). ''Speech waveform envelope cues for consonant recognition,'' J. Acoust. Soc. Am. **82**, 1152–1161.

Viemeister, N. F. (**1979**). ''Temporal modulation transfer functions based upon modulation thresholds,'' J. Acoust. Soc. Am. **66**, 1364–1380.

Viikki, O. (**2001**). ''Noise robust ASR,'' Speech Commun. **34**, 1–2.

Young, S. (**1995**). *The HTK Book* (Cambridge Research Lab, Entropic, Cambridge, England), http://htk.eng.cam.ac.uk