6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

VOWEL LANDMARK DETECTION

Andrew Wilson Howitt
Massachusetts Institute of Technology, 36-511
77 Massachusetts Avenue, Cambridge, MA 02139 USA
(617)253-5957 voice (617)258-7864 fax
howitt@mit.edu http://www.mit.edu/~ howitt/

10.21437/Eurospeech.1999-612

Abstract

Landmark based speech processing is a component of Lexical Access From Features (LAFF), a novel paradigm for feature based speech recognition. Detection and classification of landmarks is a crucial first step in a LAFF system. This work implements a Vowel Landmark Detector using a syllabic segmentation algorithm [Mermelstein 75] and examines the relative utility of its several constraints. The detector is scored against the TIMIT database, using a novel algorithm to convert the segmental transcriptions to a landmark representation for scoring. The results show that substantial improvement in performance can be gained by modifying the frequency range for peak detection. An additional advantage of this modification is that post processing to remove fricative peaks is no longer necessary, which substantially simplifies the algorithm.

Introduction

The primary motivation for this work is to detect Vowel landmarks as part of the front end of a LAFF speech recognition system. LAFF [10], [11] is a knowledge based approach to speech recognition, in which landmarks (indicating vowels, consonants, or glides) are detected in the speech signal, and phonetic features are detected and attached to the landmarks. Landmark detectors for consonants [8] and glides [12] have already been developed, leaving only vowel landmarks yet to be done.

In addition, there are many other uses for automatic syllable detection. Among them are visual speech aids for the hearing impaired, database labeling aids, tools for perceptual studies, and automatic detection of rate of articulation.

In many cases, syllable detection is an easy task. A well pronounced vowel, between well pronounced obstruent consonants, is readily detected by a simple algorithm. The challenges arise when (a) the vowel is reduced or devoiced, (b) the nearby consonants do not provide clear boundaries, because they are sonorants, semivowels, or elided, or (c) two or more vowels appear in sequence.

For examples of these phenomena, see figure 1. This is TIMIT utterance sx9 by talker mdac2, the sentence “Where were you while we were away?” A human listener can identify the distinct syllables in this sentence (using knowledge of language), but any automatic algorithm will have a very difficult time detecting separate vowel landmarks.

Most of the early work in speech recognition used knowledge based methods, attempting to explicitly incorporate speech knowledge. The expectation was that the insights of acoustics, speech production, and linguistic phenomena would aid in achieving good performance over a wide range of speech variations.

Statistically based recognition methods have largely supplanted knowledge based methods over the course of the last two decades. When the input data consist of clear, read speech, statistically based methods may achieve reasonably good performance, as shown some time ago by the results of the ARPA SUR project [5]. Only recently have statistically based systems been tested on casual, spontaneous speech, where they perform very poorly [7].

Knowledge based methods, if they can be made to work well, should be more robust against the variability in spontaneous speech, because spontaneous speech is constrained by linguistic rules that a knowledge based approach can incorporate.

After a review of algorithms for automatic syllable detection in published literature, the detector by Mermelstein [9] was selected as a starting point. The main reason for this choice is the popularity of the algorithm, which has been used in several recent projects on knowledge based speech recognition systems [2], [1]. It uses a unique recursive technique to find syllable peaks and boundaries, which reflects the effect of context by comparing the dips and peaks to their immediate surroundings. It is this recursive segmentation that appears to account for the algorithm's popularity.

This paper will describe the work on developing a Vowel Landmark Detector (VLD) based on Mermelstein's technique, and characterizing its behavior and performance. Future work will explore modifications and additional algorithms.

Algorithm issues

Mermelstein's algorithm tracks sound intensity over time, and looks for peaks and dips in the intensity track. He uses energy in a fairly broad band (500 Hz to 4 kHz, with 12 dB/octave rolloff outside that band).

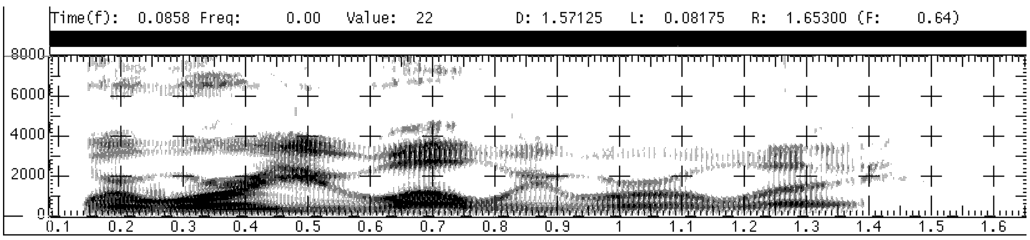

Figure 1: Example from the TIMIT database "Where were you while we were away?"

Peaks and dips are detected using a recursive convex hull algorithm [9]. The convex hull is computed by traversing the profile from its endpoints inward towards its maximum, maintaining intermediate maxima along the way. The deepest dip is compared to a threshold parameter (2 dB). If it is deeper than the threshold, the dip is accepted as a boundary, and the process recurses on the two segments thus generated. If not, the recursion ends, and the maximum is accepted as a syllabic nucleus (vowel landmark).

Mermelstein's algorithm includes durational and absolute level constraints as well. Syllabic segments are required to be at least 80 ms long (dip to dip, presumably). Syllabic peaks are required to be no more than 25 dB below the overall intensity peak. If either of these requirements is not met, the recursion terminates.

In order to prevent detection of fricative regions, the peaks must be post processed to remove peaks which appear fricative. For our implementation, we measure short time zero crossing rate (ZCR) and subject it to a threshold. A threshold of about 6000 crossings/sec appears optimal, but performance is not very sensitive to this value.

Mermelstein used two male talkers to generate slow read speech, 11 sentences each (half of which included only monosyllabic words), for a total of 22 utterances and 418 syllable tokens. While this was adequate to demonstrate the algorithm's utility, this research needs a more comprehensive database.

Among the questions we wish to investigate are: what is the relative utility of the several constraints (is one more useful than another), what is the relative sensitivity of the thresholds (how accurate do they have to be), and how do the optimal values change when testing is done on a comprehensive database, rather than the small and limited database used by Mermelstein?

The scoring problem

Any speech recognition system can only be evaluated by scoring its output against some sort of transcription of the input speech signal. If the transcription of an utterance were a straightforward, unambiguous representation of the information in the speech signal, transcribed databases would be easy to use for testing, but this is not the case.

Almost all speech databases include a *lexical* transcription, either orthographic or phonemic in nature, which represents the words being spoken. The syllable structure of a lexical transcription is generally simple and unambiguous (at least, the number and locations of syllabic nuclei, which is the focus of this research). Lexical transcriptions are also relatively easy to generate, which helps to account for their popularity.

Unfortunately, there are many cases where a lexical transcription is not a reliable indicator of the acoustic representation of syllable structure. For complete speech recognition systems (whose output is lexical in nature), this may not be a major problem. But feature extraction systems (like a Vowel Landmark Detector) attempt to represent the information in the acoustic signal without reference to a lexicon, and differences between the acoustic information and the transcription are a difficult problem to deal with.

Some speech databases include a *phonetic* transcription, which is usually a string of phones, representing the speech sounds manifested in the speech signal. Phonetic transcriptions are almost always time aligned, which helps in the scoring process.

One problem with phonetic transcriptions is that they are time consuming to generate. A more serious problem is that phonetic transcriptions are often not unique or unambiguous. A phonetic transcription imposes categorical decisions on acoustic information that varies across a continuum. How these decisions should be made is far from clear. Also, phonetic transcriptions are vulnerable to errors, because there is no easy way to check for consistency.

Mermelstein used a lexical representation (orthography) for scoring, which was not unreasonable in view of the slow, careful production of the speech being analyzed. However, he acknowledged its shortcomings, particularly its inability to represent contractions (typically unstressed syllables adjacent to stressed syllables).

	Train					Test				
	Tokens	Detect	Insert	Delete	TER	Tokens	Detect	Insert	Delete	TER
broadband	7585	76.2%	2.40%	23.8%	26.2%	4404	77.7%	2.61%	22.3%	24.9%
F1 range	7585	88.4%	1.82%	11.6%	13.4%	4404	87.1%	1.73%	12.9%	14.6%
	without post processing for fricative detection									
broadband	7585	87.8%	21.8%	12.2%	34.0%	4404	87.8%	19.7%	12.2%	31.9%
F1 range	7585	88.4%	1.96%	11.6%	13.6%	4404	87.1%	1.88%	12.9%	14.8%

Table 1: Scores by frequency range, with and without fricative detection. The "broadband" condition is Mermelstein's original frequency range (500 Hz - 4 kHz), and the "F1" range is 0 - 650 Hz.

Scoring paradigm

In an effort to achieve accurate scoring, we introduce the notion of Reference Vowel Landmarks (RVLMs) which are derived from the phonetic transcription, and which may be compared to Detected Vowel Landmarks (DVLMs) which are the output of the VLD.

RVLMs are Vowel landmarks, just like the output of the VLD, but they are derived from the database transcription. Different databases have different transcription formats (orthography, aligned phonetic segments, landmarks) which would require different scoring techniques if they were used directly. By converting them to RVLMs, we can use the same scoring technique across different databases, permitting direct comparison of the results.

The RVLM format allows for optional matching. Every RVLM must match at least one DVLM to avoid a deletion error. The data format allows RVLMs to match optionally more than one DVLM. This enables the scoring procedure to avoid penalizing marginal or ambiguous cases.

At present, we allow optional landmarks only for directly abutting vowels, which may be subject to coalescence or merging. These are the places where the database transcriptions seem to be questionable or inconsistent. Other phenomena, such as vowel deletion or epenthetic insertion, ought to be represented in the acoustic transcription, and we will trust that the transcription is accurate enough for experiments.

The TIMIT database [6] is a database of read sentences recorded in quiet. It includes separate training and test data sets, by hundreds of talkers, covering eight regional dialects of American English, all with aligned phonetic transcriptions.

The phonetic transcriptions are converted to RVLMs using a syllable parsing program called TSYLB [3]. Its syllable grammar operates on a string of phones, upon which it places syllable structure. The nucleus of each syllable generated by TSYLB is taken as the reference landmark, and its time is halfway between the start and end times for the corresponding vowel segment. Multiple vowels in sequence are represented by one RVLM, whose time is halfway between the start and end times for the sequence, and which can optionally match as many DVLMs as vowels in the sequence.

For the experiments reported in this paper, all sentences in the TIMIT database whose numbers end in 8 or 9 were selected (male and female, all dialects). The aligned phonetic transcriptions for these utterances were examined, and a few (about 1%) were discarded because the transcriptions were inconsistent or not parseable. Results for the training set were 619 utterances and 7585 syllable tokens, and for the test set, 373 utterances and 4404 syllable tokens.

Experimental results

The fundamental value used to characterize performance is Token Error Rate (TER), which is the sum of insertions and deletions as a percentage of syllable tokens. In all cases, performance is almost the same on training and test data, indicating that the training set includes adequate representation of speech phenomena in this database.

The first experiment attempted to reproduce Mermelstein's algorithm exactly as described in [9]. Experimental results are shown in the first row of table 1. Overall performance was 26.2% TER, substantially worse than Mermelstein reported (9.5% TER, 6.9% deleted, 2.6% inserted). Part of this difference may be due to the more comprehensive data set (rapid speech by many talkers of both genders and diverse dialects). There may also be differences in the details of implementation.

It was felt that the broadband intensity measure was not optimal. The definition of vowel landmarks specifies that they should be located around peaks in energy in the region of F1. If so, the performance should improve when the intensity is measured in a band around F1, nominally about 300 to 900 Hz.

To investigate the effect of this band, the upper and lower band edges and their rolloff values were varied independently. The optimal frequency band (0 to 650 Hz) does indeed delineate the region where F1 is likely to be found. Performance on this band is 14.6% TER (second row of table 1).

Observing that fricative regions are much less likely to be detected when using the F1 band, we inves-

	Train					Test				
	Tokens	Detect	Insert	Delete	TER	Tokens	Detect	Insert	Delete	TER
Overall	7585	88.4%	1.96%	11.6%	13.6%	4404	87.1%	1.88%	12.9%	14.8%
Tense	3168	91.7%	3.22%	8.27%	11.5%	1838	89.6%	2.99%	10.4%	13.4%
Lax	1704	92.8%	0.59%	7.22%	7.81%	1001	91.7%	1.10%	8.29%	9.39%
Schwa	2458	81.9%	1.42%	18.1%	19.5%	1446	81.2%	1.11%	18.8%	19.9%
Sonor	255	80.4%	0.784%	19.6%	20.4%	119	83.2%	0.00%	16.8%	16.8%

Table 2: Scores by vowel stress. In general, less stressed vowels are more difficult to detect. The exception is lax vowels, which are easier to detect in context (because they are always followed by consonants).

tigated performance without fricative detection. As expected, performance is very poor when using broadband intensity (third row of table 1, 34% TER!), but much better when using F1 intensity (fourth row of table 1, 14.8% TER). It appears that fricative detection offers essentially no performance gain when using the F1 frequency band. Therefore, all subsequent experiments were done using the F1 band without fricative detection.

Details of the detector's performance by vowel stress are shown in table 2. As one might expect, full vowels (tense and lax) are easier to detect than schwas or syllabic sonorants. Lax vowels are the most easily detected, even more so than tense vowels. This may result from the fact that lax vowels must be followed by consonants, whereas tense vowels may not be. The consonants provide clear boundaries for segmentation, making detection easier.

Conclusion

Substantial improvement in performance over Mermelstein's original algorithm can be gained by modifying the frequency range for peak detection, to focus on the first formant. An additional advantage of this modification is that post processing to remove fricative peaks is no longer necessary, which substantially simplifies the algorithm.

With this scoring technique in hand, more additions and modifications to the detection algorithm can be explored. Stevens [11] has suggested placing vowel landmarks at peaks of F1 frequency or amplitude, which would require a formant tracker. Other researchers have used direct measures of formant presence in vowels [4] and fuzzy logic techniques for combination of multiple acoustic cues [1]. The data presented here will serve as a starting point for more comprehensive evaluation of methods for vowel landmark detection.

References

[1] Bitar, Nabil N. *Acoustic Analysis and Modeling of Speech Based on Phonetic Features.* Ph. D. thesis, Boston University, 1997.

[2] Espy-Wilson, Carol Y. *An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels.* Ph. D. thesis, Massachusetts Institute of Technology, 1987.

[3] Fisher, W. Program TSYLB (version 2 revision 1.1), NIST, 7 August 1996.

[4] Hermes, D. J. "Vowel onset detection." JASA 87(2):866-873, February 1990.

[5] Klatt, D. H. "Review of the ARPA Speech Understanding Project." JASA 62(6):1345-1366, December 1977.

[6] Lamel, L. et al. "Speech Database Development: Design and Analysis of the Acoustic–Phonetic Corpus." Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, 100-109.

[7] Lippmann, R. P. "Speech recognition by machines and humans." Speech Communication 22(1):1-15, 1997.

[8] Liu, Sharlene A. "Landmark detection for distinctive feature-based speech recognition." JASA 100(5):3417-3430, November 1996.

[9] Mermelstein, P. "Automatic segmentation of speech into syllabic units." JASA 58(4):880-883, October 1975.

[10] Stevens, K. N. "Models of Phonetic Recognition II: An Approach to Feature–based Recognition." Paper 5.5, Montreal Symposion on Speech Recogntion, Proceedings. McGill University, 1986.

[11] Stevens, K. N. "Phonetic Features and Lexical Access." Paper 10, Second Symposium on Advanced Man–Machine Interface Through Spoken Lanuage. Makaha, Hawaii, 1988.

[12] Sun, Walter. "Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition." Masters thesis, MIT, November 1996.