A MULTI-STAGE PROCEDURE FOR IDENTIFICAITON OF ACOUSTIC MICRO SEGMENTS IN STOP+VOWEL+STOP SYLLABLES.

Terrance. M. Nearey and Michael Kiefte
University of Alberta, Edmonton, Canada T6G 2E7

ABSTRACT

This paper describes an efficient multistage algorithm for the parsing of stop-vowel-stop syllables into acoustic microsegments typically used for scientific measurement of speech data and as well as in some ASR applications. The methods adopted illustate one way in which expert speech knowledge and powerful statistical pattern-recogniton can be usefully combined to provide robust and intuitively satisfying solutions.

OVERVIEW

Six microsegments of interest were established to characterize /CVk/ syllables, where the initial consonants ranged over /p, t, k, b, d, g/ and where the vowels ranged (in the test sets) over the 15 Canadian English vowels and diphthongs. The microsements are labeled *M*1 to *M*6 and are listed in Table 1.

Table 1. Microsegment categories for segmentation algorithm.

Microsegment	*description*
*M*1	initial (C) silence
*M*2	initial voicebar
*M*3	initial C release
*M*4	vowel
*M*5	final C silence
*M*6	final C burst

The object of the algorithims described is to produce reliable estimates for the beginings of the microsegments *M*2, *M*3, *M*4 and *M*5.

The first stage uses statistics from a set of hand marked segments to provide a preliminary "fuzzy categorization" (in the form of a posteriori probabilities, henceforth APP scores) for each microsegment type and for each 20 ms rectangular window of speech advanced through the signal in 1 ms frames. Statistics are calculated from feature vectors for each frame which consist of six readily calculated properties. A second stage applies a Viterbi search method adapted from a continuously variable duration hidden (semi-)Markov model to bracket regions within which cursors can be placed. A third stage uses various heuristic measures to set the exact locations of specific cursors within the bracketed regions.

STAGE 1

The feature vectors for each frame consist of mean absolute amplitudes of 10 ms sections before and after the frame centers for 1) the original signal, 2) a high pass ($1/1\text{-}z^{-1}$) signal and 3) a band-pass ($1/1\text{-}z^{-2}$) filtered signal. The maximum mean absolute amplitude over all frames for the entire syllable was also included as a normalization measure.

Means, vectors and covarinace matrices were calculated using these features for the basic signal types shown in Table 2.

Table 2. Basic signal types defined by feature vectors.

Signal Type	*Description*
*D*1	silence
D2a	voice bar onset
D2b	voice bar cont.
D3a	C1 burst onset
D3b	C1 fric./asp.
D4a	V onset
D4b	V continuation
D4c	V end
*D*5 (=*D*1)	final /-k/ silence
D6a (=*D3a*)	final /-k/ burst
D6b (=*D3b*)	final /-k/ fric./asp.

The numbers associated with each distribution indicate the microsegment from Table 1 associated with each distribution. Voice bar and consonant release are each associated with two distributions. Those labeled *a* (e.g. *D2a*) correspond to the onset part of the microsegment and those labeled with *b* correspond to continuation of the microsegment type. The distribution type "vowel" is associated with three distributions. Training samples for onset type *a* segments (as well as *D4c*) were centered on hand-marked cursors that delineated clearly acoustic boundaries, so that the features associated with the first and second 10 ms parts of the frame would be expected to differ significantly. Those for the continuation type *b* distibutions were more likely to be homogeneous. Training data for the *b* type distributions were extracted from the centers of hand-marked microsegment types of at least 30 ms duration to avoid the onset and offset of the microsegments) Only distributions *D*1 through *D4a,b,c* were explicitly trained, since hand marked cursors did not exist for portions of the signal following *M*4. Distributions *D*5 and *D6a,b* were effectively "tied" to the corresponding distributions for the variable initial stops.

APPs were calculated for distributions *D*1 to *D4a,b,c*. Trials with a training set of 703 syllables indicated that these types could be identified with a reasonably high rate from the hand-marked data.

Running plots of the APP scores for stimuli were examined and indicated that the scores in question would serve as a useful basis for parsing the signal, since an appropriate distribution usually showed the highest APP for most of the duration of the target microsegments. (Appropriate distributions include either of the *a,b* pairs of microsegments associated with onset and continuation type distributions for reasons discussed below.)

STAGE 2

A modified continuously variable duration semi-HMM (CVDHMM) [1] was employed to group the frames of preliminary types *D*1-*D*6. The states of the CVDHMM are outlined in table 3 with their associated signal type. Possible transitions of the CVDHMM are shown in Figure 1.

The modifications from the CVDHMM of [1] consisted of three main simplifications involving "engineered" rather than optimized esteimates of parameter values. First, distribution parameters of Table 1 were fixed in advcance as described above and were not reestimated. Second, state transitions probabilities were determined *a priori*: the probability of the each exit path was set to the reciprocal of total number of such paths for that state. Third, although Gaussian state-duration distributions were originally investigated, it was found that they had little effect on the results and a simpler method ignoring state-durations entirely was adopted. This is equivalent to assuming a uniform distribution of all feasible durations (e.g. 1 to 1000 ms) for all states in a CVDHMM framework.

Table 3. CVDHMM states.

State	*Description*
*S*1 (*D*1)	silence
*S*2 (*D2a*)	C1 voicebar
*S*3 (*D2b*)	C1 voicebar cont.
*S*4 (*D3a*)	C1 burst
*S*5 (*D3b*)	C1 asp./fric.
*S*6 (*D4a*)	V onset
*S*7 (*D4b*)	V continuation
*S*8 (*D4c*)	V end
*S*9 (*D*5)	C2 silence
*S*10 (*D6a*)	C2 burst
*S*11 (*D6b*)	C2 asp./fric.

Preliminary observation of the results from the CVDHMM coarse segmentation showed that state changes were nearly always limited to points of large changes in the Stage 1 APP scores and that a substantial reduction of the search space for stage could be accomplished by using a simple preliminary thresholding of changes in APP scores. A measure of change in APP scores for each of the distribution types was calculated as

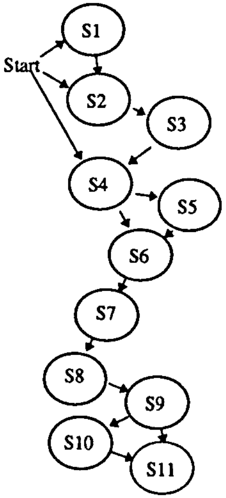
$$\delta_i = \sum_{k=1}^{n} \text{abs}(\text{W}(\text{med}_3(m_{k,i-2}), \text{med}_3(m_{k,i+1})))$$

where $m_{k,i}$ represents the six-point median filtered APP score for category k at time i, $\text{med}_3(x)$ indicates the median of three arguments centered at point x and the function W is calculated as

$$W(x,y) = \frac{x}{1-\text{abs}(x-0.5)} - \frac{y}{1-\text{abs}(y-0.5)}$$

which weights the function in favor of high APPs. The threshhold was set at $\delta_i > 0.4$ where CVDHMM transitions were allowed to occur in the Viterbi search algorithm.

Figure 1. Non null state transitions for CVDHMM.



STAGE 3

Although the parsing provided by the CVDHMM was found to correctly bracket most of the events of interest, a final heuristic post-processing stage was used to fine-tune the placement of selected segment boundaries. The start of *M*2 was determined as

$$C_{vb} = \arg\max_{t}\left(\frac{\text{L}'(t,1)\text{R}(t,1)}{\text{R}'(t,1)\text{L}(t,1)+0.1}\right)$$

where R(i,1) is the average absolute value of the waveform in a 1 ms window to the right of point i, and L(i,1) is a similar measure taken from the left. R′(i,1) and L′(i,1) are the same measures taken from the low-passed filtered feature vector. t is limited to samples bounded by *S*2 and the first 30 ms of *S*3.

*M*4 was then located by exactly the same method with the values of t limited to the samples bounded by *S*6 and the first 30 ms of *S*7, with the exception that interval of averaging was 3 ms.

The following measure was then used to fine tune the start position of *M*3:

$$C_{b} = \arg\max_{t}\left(\frac{\text{R}(t,3)+\text{R}'(t,3)+A_{\text{min}}}{\text{L}(t,3)+\text{L}'(t,3)+A_{\text{min}}}\right)$$

where the integrated absolut value of the waveform is taken from 3 ms windows and A_{min} is the minimum mean absolute amplitude of the original signal. The search was restricted to the samples between the estimated start of *M*2 and the sample 30 ms following the start of *S*7.

The final position of *M*4 was taken as the window at which the product of the probability density function and the APP score was a maximum for *D*4*c* within the bounds of *S*8.

Informal evaluation of the algorithm indicates excellent agreement with human operator jugments in most cases. (More formal evaluations are in preparation and will be presented at the conference). The procedure has potential for use in semi-automated data collection for descriptive linguistic and speech database applications.

ACKNOWLEDGEMENTS

We would like to thank Tom Weltz and Terry Baxter for their programming skills. The research for this paper was supported by AGT.

REFERENCES

[1] Ljolje, A., & Levenson, S. E. (1991). "Development of an acoustic-phonetic hidden Markov model for continuous speech recognition", *I.E.E.E. Transactions on Signal Processing*, vol. 39, pp. 29-39.