Speech Communication 2 (1983) 190–192
North-Holland

AUTOMATIC ALIGNMENT OF NATURAL SPEECH WITH A CORRESPONDING TRANSCRIPTION

Matthew LENNIG

Bell-Northern Research and INRS-Télécommunications, 3 Place du Commerce, Ile des Soeurs, Montréal, Québec H3E 1H6, Canada

Received 18 March 1983

The goal of the present research is to devise an algorithm for the automatic alignment of a speech signal with its phonetic transcription. We assume that both the signal and its phonetic transcription are known. What is required is to locate the phone 'boundaries' in the signal, with a precision comparable to that of phoneticians performing the same task.

The theoretical question posed by automatic alignment research is how to represent the knowledge used by a phonetician to segment a spectrographic representation of a known utterance. If no transcription of the utterance is available, both human and computer segmentation are extremely difficult. Although certain types of boundaries can be located reliably (e.g., vowel/fricative), other types present serious difficulties (e.g., vowel/liquid). Once the transcription is known, however, a phonetician has little difficulty in assigning a segmentation and labelling to a spectrogram or other suitable representation of the speech signal. Although certain ambiguities in segment boundary location cannot be resolved, conventions for dealing with most of these cases can be established (cf. Peterson and Lehiste 1960).

The practical importance of a reliable phonetic alignment algorithm is to facilitate the segmentation and labelling of large amounts of natural speech for use in large scale statistical studies of speech variation. Such studies are likely to provide basic speech knowledge needed for successful speaker-independent recognition of continuous speech. Also, the shortcomings of a particular automatic alignment algorithm may themselves give clues to the inadequacies of current speech models and representations.

A published alignment technique due to Wagner (1981) may be characterized as partially 'bottom up' since it requires a prior division of the waveform into a sequence of acoustic segments with the labels voiced, voiceless, and silence. The analysis is then refined, yielding an assignment of phone labels to individual frames. The present paper explores a 'top-down' approach in which the natural speech is time-warped against a synthesized version of the same text, generated from the given transcription via synthesis-by-rule. The warp path is used to map phone boundaries from the synthetic reference onto the natural speech, inducing on it a phonetic segmentation and labelling. Phone boundary locations in the synthetic reference are determined by a rule-based parsing of the parameter stream used to drive the synthesizer.

Rule-based segmentation of the synthetic reference

Initial experiments in the segmentation and labelling of natural speech using the MITalk-79 [1] text-to-speech system to produce a segmented reference utterance revealed that the nominal segmentation used by MITalk was inappropriate. The segment size (phonemic) was considered too large to permit precise evaluation of the alignment technique. Boundary placement, though not incorrect, required rationalization. For these reasons, and in order to allow flexibility in determining segmentation rules for the synthetic reference, a rule-based segmenter was designed to segment the stream of

[1] The MITalk text-to-speech system is used with permission of MIT.

0167-6393/83/$3.00 © 1983, Elsevier Science Publishers B.V. (North-Holland)

parameter frames used to drive the Klatt synthesizer.

The main advantage of using the rule-based segmenter is that it allows the segment size, segment inventory, and segmentation criteria to be independent of the synthesis-by-rule system used for alignment. For the present study, a level of transcription was chosen to reflect certain phonetic events not present in the transcription used by MITalk, such as the distinction between stop closure and stop burst. Affrication was considered, by convention, to be part of the stop burst. If aspiration followed the stop burst it was considered to be part of the vowel.

The rule-based segmenter requires that each segmentation criterion be stated explicitly. The segmenter, implemented in LISP, employs the following fourteen phonetic categories, into which all phonetic segments are classified: vowel, liquid, glide, nasal consonant, voiced nonsibilant fricative, voiceless nonsibilant fricative, voiced sibilant, voiceless sibilant, voiced stop closure, voiced stop burst, voiceless stop closure, voiceless stop burst, voiceless aspirate, silence. Rules specify which events must occur in the stream of parameter frames in order to make a transition from one phonetic category to another. For example, a boundary between a vowel and a nasal consonant requires a sudden large change in formant bandwidths.

In the mode described above, the segmenter is a finite state machine. However, because certain kinds of boundaries are context dependent, e.g., those between vowels, liquids, and glides, segmentation decisions are deferred until the end of such a sequence is reached, at which point rules based on formant extrema, medians, and rates of change are used to complete the segmentation. Due to the regularity of synthetic speech and because no measurement error is introduced in the estimation of acoustic parameters, it has been possible to develop segmenter rules which produce precise, reliable segmentation of the synthetic reference speech.

Alignment experiments

The following sentence was pronounced by four male speakers and synthesized using the MITalk text-to-speech system:

The sink is the thing in which we pile dishes.

Three of the speakers were native speakers of Montreal English; the fourth was a native speaker of New York English. The naturally produced sentences were lowpass filtered at 4.4 kHz and sampled at 10 kHz. The endpoints of each sentence were manually determined. The synthetic sentence was preprocessed to yield a mel-frequency cepstrum every 5.0 ms. The naturally produced sentences were processed similarly, except that the frame advance for each speaker was chosen between 4.7 and 6.4 ms so as to yield approximately the same number of speech frames as in the synthetic reference (454 frames). Unequal numbers of frames are undesirable in the time warping procedure because slope constraints and slope penalties are used to make the warp path tend toward a 45 degree line.

The decimated-grid, symmetric time warping algorithm proposed by Mermelstein (1978) was used, in which grid point (i, j) is accessible from points $(i-1, j-1)$, $(i-2, j)$, and $(i, j-2)$. The penalty for a vertical or horizontal step is to multiply the local distance at (i, j) by 1.5. Two different local slope constraints were tried: unconstrained and constrained slope. The constrained slope algorithm permits a maximum of one consecutive vertical or horizontal step. In a third trial, the first speaker's utterance was hand segmented and used as a reference in conjunction with the constrained slope algorithm to induce segmentations on the remaining three naturally produced sentences.

Results and discussion

Segmentation and labelling induced on the natural speech was inspected by viewing spectrograms annotated with this information. Each automatically determined segmentation boundary was subjectively classified as correct or incorrect. Subjective scoring was preferred because certain segment boundaries, such as the endpoints of a stop burst, are more precisely determined by the speech signal, while others, e.g., the boundary between a vowel and a liquid or glide, may be farther away from a prescribed norm and still be considered correct.

Table 1
Error rates for segment boundary location using constrained and unconstrained time warping algorithm

	Synthetic reference		Natural reference
	unconstrained	constrained	constrained
sonorant/ sonorant	26/32 (81%)	10/32 (31%)	4/24 (24%)
nonsonorant/ nonsonorant	12/20 (60%)	8/20 (40%)	2/15 (13%)
sonorant/ nonsonorant	13/72 (18%)	9/72 (13%)	6/54 (11%)
Total	51/124 (41%)	27/124 (22%)	12/93 (13%)

Table 1 displays segmentation errors according to three boundary types: boundaries between sonorant segments (vowels, liquids, nasals, glides), boundaries between nonsonorant segments, and boundaries between a sonorant and a nonsonorant segment, in either order. As expected, boundaries between similar segment types give rise to higher error rates. Segmentation performance for all boundary categories is significantly better when a slope constraint is employed and significantly better using a natural speech reference as compared with a synthetic reference. Even using a natural reference, however, the 13% error rate obtained is unsatisfactory for the application cited above.

The advantage of the constrained time warping algorithm over the unconstrained algorithm appears most saliently in the detection of sonorant/sonorant boundaries. This may be because interspeaker spectral differences in sonorants overshadow spectral differences between different sonorant segments, leaving segment duration as the only reliable alignment criterion. Use of a speaker normalization for glottal source spectrum shape may improve the accuracy of sonorant/sonorant boundary localization.

One finding obscured by Table 1 is that stop burst localization contributes significantly to the error rate. In the constrained, synthetic trial, for example, seven of the eight nonsonorant/nonsonorant errors occur on closure/burst boundaries. Five of the nine sonorant/nonsonorant errors occur on burst/vowel boundaries. Similar results hold for the other trials. Often, what is observed on the annotated spectrogram is an erroneous localization of the stop burst somewhere in the middle of the stop closure, temporally disjunct from the actual burst. Such errors can be explained by the relatively small distance penalty incurred by burst misplacement, due to the segment's short duration. The problem should be viewed as an inadequacy of the time warping algorithm as currently formulated.

References

P. Mermelstein (1978), "Recognition of monosyllabic words in continuous sentences using composite word templates", *Proc. 1978 IEEE Int. Conf. Acoust. Speech Signal Process.*

G. Peterson and I. Lehiste (1960), "Duration of syllable nuclei in English", *J. Acoust. Soc. Am.*, Vol. 32, No. 6, pp. 693–703.

M. Wagner (1981), "Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms", *Proc. 1981 IEEE Int. Conf. Acoust. Speech Signal Process.*