RECOGNITION OF MONOSYLLABIC WORDS IN CONTINUOUS SENTENCES USING COMPOSITE WORD TEMPLATES

Paul Mermelstein

Bell-Northern Research and INRS-Telecommunications
Montreal, Quebec, Canada H3E 1H6

Abstract

As a preliminary step in syllable recognition in continuous speech, differences in acoustic forms were studied when these are induced by variation in the syntactic role of the word in the sentences. A modified dynamic programming algorithm is presented that allows building up of reference information from a speaker's productions in the face of such variations. The limited data base studied included 169 productions occurring in 57 sentences and composed of 52 different monosyllabic CVC words spoken by two male speakers. When each speaker's productions were tested against reference data built up from different tokens of his productions, correct first choice recognition was attained for 83% and 90% of the words, respectively. A similarity metric that takes into account the detailed acoustic variation that each speech sound may undergo without loss of identifiability can be expected to improve on these results.

Introduction

To recognize a word in continuous speech one needs to develop a measure of similarity that reliably indicates the reference word most similar to an unknown word. To verify that the unknown is a token belonging to some reference class, a threshold on the measure of similarity must be satisfied. The success of both procedures strongly depends on the selection of a measure that adequately reflects perceptual criteria of similarity. As an initial step in developing such a measure, we have evaluated the recognition of monosyllabic words based on similarity with reference forms generated from other occurrences of the same word by the same speaker. A modified dynamic programming algorithm is presented that allows the generation of a composite reference pattern from two tokens of the same word. The composite is the result of averaging the acoustic information over time points mapped onto each other by the path of maximum similarity resulting from the dynamic comparison. To include information from additional tokens in the reference form, we apply the same averaging operation to new tokens and the previously found composites.

The recognition techniques discussed here are oriented towards the comparison of syllable-sized segments. The comparisons that need to be carried out depend on the organization and scope of the speech recognition system employing the techniques. Through detailed context-independent acoustic analysis one may narrow the set of hypotheses applicable to a given speech segment to a small number of alternatives. In such situations, matching an unknown only to those syllable-sized segments that appear in a dictionary of syllabic forms admitted by the vocabulary can be effective in further reducing the number of possible phonetic interpretations. Our results, so far, pertain to monosyllabic consonant-vowel-consonant words. Instead of a detailed acoustic analysis, we have used vowel-specific templates to reduce the number of necessary syllabic comparisons. In the learning phase templates are generated for all distinct CVC forms as well as one for each vowel independent of the consonantal context. In the recognition phase we first compare the unknown syllable with all vowel templates and only the syllabic reference forms that correspond to the vowels found most similar are subjected to further detailed comparisons.

The comparison operation and its application to template generation

The comparison operation used employs a modified form of the dynamic programming algorithm first applied to spoken words by Velichko and Zagoruyko (1970) and subsequently modified by Bridle and Brown (1974) and Itakura (1975). An intermediate result in the computation of the minimum distance or maximum similarity between two tokens is that path which maps the time-coordinates of the two tokens onto each other so as to result in maximal overall similarity. To generate composite templates, that same path is used to determine the pairs of time-coordinates which link the respective acoustic variables to be averaged. The result is a path in acoustic space whose similarity to the paths corresponding to the individual tokens is roughly maximized. Due to the nonlinear nature of the similarity computation, however, this result does not precisely correspond to that which would be obtained by simultaneous minimization of the distance between the composite and each of the tokens.

To create a time-warping algorithm suitable for template generation, modifications were introduced to the previously used dynamic programming

CH1285-6/78/0000-0708$00.75@1978IEEE

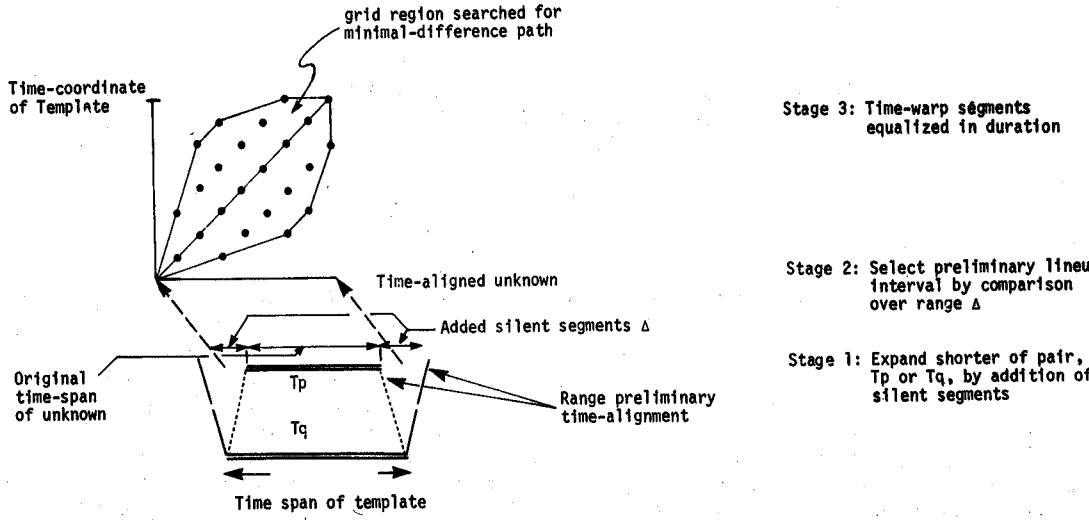

Fig. 1 - Preliminary Time-Alignment and Nonlinear Time-Warping Operations

methods to make the computed distance between a pair of tokens a symmetric function independent of the order of the pair. This modification tends to reduce, although it does not completely eliminate, the dependence of any composite template generated from a larger group of tokens on the order in which the tokens are combined to build up the templates.

A second modification introduced is applicable mainly to syllabic segments. To equalize the time spans of the tokens to be compared, a preliminary time alignment is added that first extends the shorter segment both at its beginning and end by an interval of silence equal in duration to the original time difference between the tokens. The procedure is illustrated as stage 1 of Figure 1. It next lines up the originally longer token with the one newly extended through a time shift that maximizes the integrated similarity measure between them (stage 2). This procedure has been found more effective than linear time alignment of the two tokens to an intermediate duration because it does not introduce changes in the duration of many consonantal segments where such durations could carry significant phonetic information.

The nonlinear segment of the time alignment procedure (stage 3) attemps to increase the overall similarity computed for the two tokens by exploring off-diagonal paths in the space of the two time-coordinates. Let t = 1, ..., T be integer time values spanning the two tokens, one possibly lengthened through the addition of silence, and $\{\underline{a}(t)\}$, $\{\underline{b}(t)\}$ the corresponding acoustic representations of the tokens to be compared. For each t we explore points $\{x(t, n_t)\}$, $|n_t| < N_t$, along a line normal to the diagonal path represented by $\{x(t,0)\}$. A point on the grid $x(t, n_t)$ is accessible from some previously considered point $x(t-1, n_{t-1})$ if it satisfies a maximal time-warp slope criterion $|n_t - n_{t-1}| < 1$. Then the integrated minimal distance from the origin to x is given by

$$D(t, n_t) = \min_{n_{t-1}} \left[D(t-1, n_{t-1}) + d_x[\underline{a}(t-n_t), \underline{b}(t+n_t)]. \ \nu(n_t - n_{t-1}) \right]$$

where $d_x[\underline{a}, \underline{b}]$ is a local difference function and ν is a penalty function set to 1.5 for $|n_t - n_{t-1}| = 1$, and 0 otherwise. Define the predecessor of $x(t, n_t)$ as $x(t, \eta_{t-1})$ where η_{t-1} is that n_{t-1} which minimizes $D(t, n_t)$. An additional time-warp slope criterion is imposed between $x(t, n_t)$ and the predecessor of its predecessor $x(t-2, \mu_{t-2})$, namely $|n_t - \mu_{t-2}| < 2$ in order to restrict any one time frame to participation in at most two local comparisons. Since all paths end in x(T, 0), the total distance of the minimum distance path is D(T, 0).

Data collection

Our initial experiments focused on variability induced in words of the same speaker by variation in the surrounding words and by syntactic position. Since the storage of a separate reference template corresponding to each different syntactic environment would significantly increase the memory requirements in any practical implementation, we felt it important to assess the viability of a procedure that combined tokens irrespective of syntactic position in the generation of reference templates. The speech data examined comprised 52 monosyllabic CVC words from two male speakers. 169

tokens were collected from 57 distinct sentences for each speaker.

To achieve the required variability, we selected words that could be used as both nouns and verbs. For example, "Keep the hope at the bar" and "Bar the keep for the yell" are two sentences that allow syntactic variation but preserve the same overall intonation pattern. All the words examined carried some stress, the unstressed function words have not yet been analyzed. The target words, all CVC's, included 12 distinct vowels,/i, I, e, ɛ, æ, ɔ, ʌ, U, u, ɝ, a, o/ some of which are normally diphthongized in English. Each vowel was represented in at least 4 different words, including differences in both the prevocalic and postvocalic consonants. The consonants included simple consonants and affricates but no consonantal clusters.

Acoustic analysis

An automatic segmentation process (Mermelstein, 1975) was used to assign initial and final boundary markers to each CVC token in turn. This segmentation is based on FFT spectra spaced 12.8 ms in time. The results of the segmentation were manually verified and adjusted where required. Where final stops were released, the release bursts were not generally included in the segments further analyzed. For initial stops an attempt was made to include at least one frame prior to the release so as to retain information regarding the rapidity of the initial onset.

The spectral information within the delimited segments was converted to log energy within 20 contiguous frequency bands spaced linearly, 100 Hz apart, to 1 kHz, and logarithmically between 1 and 4 kHz. Six cepstral coefficients were determined for this mel-based frequency scale as the first six coefficients of the cosine expansion of the logarithm of the filter energies. The choice of six coefficients was dictated by the need to achieve adequate low-frequency resolution without sacrificing economy in computing and memory requirements.

Vowel analysis results

Three distinct procedures were attempted to achieve a preliminary hypothesis regarding the syllabic vowel and thereby reduce the number of subsequent CVC comparisons. We deliberately avoided techniques that required explicit indication of the time extent of the vowel segment. The first procedure attempted to identify a salient point in time that gave the best indication of the vowel spectrum. The point selected was that where the loudness function, a frequency weighted energy function, reached its maximum value for the syllable. Classification of one speaker's vowels in a six-dimensional variance-weighted vowel-space resulted in only 48% correct first-choice vowel recognition and in an average rank of 2.6 for the correct vowel hypothesis.

To attain improved stability in the selection of the vowel spectrum, we next attempted to determine a weighted time-average vowel spectrum through the use of the mean log energy (zeroeth cepstral coefficient) as the weighting factor. Vowel recognition improved to 63% and an average rank of 1.8 for the correct hypothesis.

The inherent nonstationarity of vowels in the syllabic context prevented the selection of one time frame as a best indicator of the spectral characteristic of the vowel. To overcome this problem, we generated vowel-specific syllable templates from all the productions of our speaker that included each particular vowel. The technique previously outlined for syllable template generation was used for the generation of the vowel templates as well because it is effective in building up the invariant information in the region of the syllabic nucleus and to smear out the acoustic information in the consonantal region where the tokens exhibited marked differences. Reference to vowel-based templates resulted in correct first-choice vowel selection for 82% of the tokens of speaker DZ and 69% for speaker LL. Perhaps a more important criterion for the evaluation of the vowel analysis is the number of CVC comparisons avoided by prior vowel comparison without rejection of the correct vowel with significant frequency. Use of a distance threshold equal to twice the minimum distance between the unknown and all vowel templates, resulted in less than 2% false misses for the tokens of each speaker. 74% of the CVC tokens were correctly excluded.

Syllable analysis results

The syllable comparison was applied to all reference tokens that incorporated vowels found similar to the unknown. For each unknown, a leave-one-out procedure was utilized, which required the modification of the template actually corresponding to the unknown by subtracting (in the generalized sense) the unknown from the previously formed composite. Due to the nonlinear nature of the dynamic warping process, this subtraction operation does not yield a result precisely equal to the template that would have been obtained without including the unknown in the first place. However, the resulting bias was estimated to be significant only if the time-warp map between the unknown and the template deviated significantly from the diagonal.

A minimum distance classification procedure applied to the possibly modified templates resulted in 90% and 83% correct first choice recognition for the two speakers. In each case a speaker dependent mode was utilized, i.e., unknown and reference date were produced by the same speaker. An important additional feature of our method is that when a hypothesis is rejected, the time-distributed difference function can be used to select a hypothesis that is likely to be better. In Fig. 2a we illustrate the difference function for a correct hypothesis; In Fig. 2b the difference function for the incorrect hypothesis points to a suspected error in the initial consonant.

To test the effectiveness of the use of composite templates, a further test was run in which the reference data included only individual tokens and no composite templates. To save processing time, the reference corpus was restricted to those

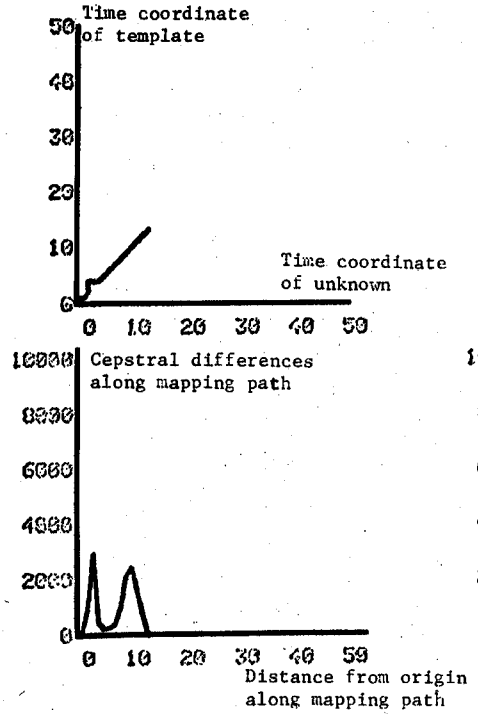

Fig. 2a.
Unknown = /bUk/
Template = /bUk/

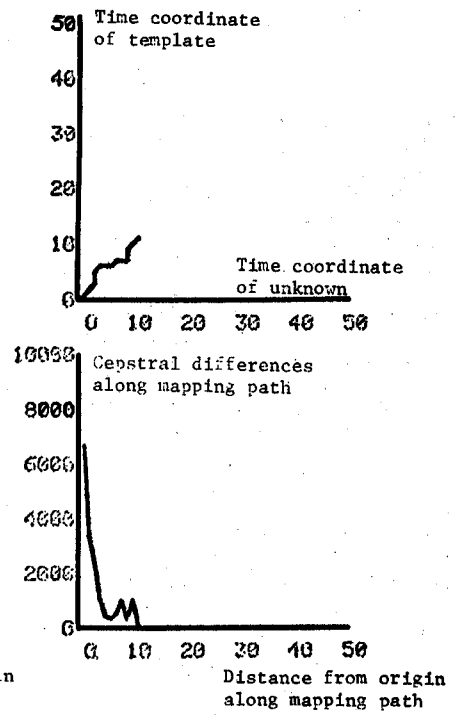

Fig. 2b.
Unknown = /hUk/
Template = /bUk/

tokens found similar through the vowel analysis technique previously described. Recognition accuracy for the syllables of speaker DZ dropped from 90% to 58%. Evidently, the use of composite templates not only reduces the storage requirements for the reference data, but also improves recognition by retaining only acoustic information consistent among the tokens corresponding to an individual word.

Conclusions

1) The use of composite word templates appears to be a promising technique for the recognition of monosyllabic words in continuous sentences. Generalizability to multisyllabic words remains to be demonstrated.

2) Dynamic warping of mel-based cepstral coefficients can significantly reduce the time variation between differences among tokens of the same word induced by syntactic variation within the utterance. The dynamic time-warping must be applied carefully in order to minimize time adjustments in the consonantal regions where such variations carry significant phonetic information.

3) Use of composite templates generated from only 2 or 3 tokens does not yield reliable estimates of the variability of the acoustic parameters pertaining to an individual word. Evaluation of the contribution of second-order parameter statistics to the improvement of recognition rates remains to be carried out.

Acknowledgements

This research was supported in part by NSF Grant MCS76-81034 and USECOM Contract MDA 904-77-C-0157 to Haskins Laboratories.

References

Bridle, J.S. and M.D. Brown (1974). An experimental Automatic Word Recognition System, JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. ASSP-23, 67-72.

Mermelstein, P. (1975). Automatic Segmentation of Speech into Syllabic Units, J. Acoust. Soc. Am. 58, 880-883.

Velichko, V.M. and N.G. Zagoruyko (1970). Automatic Recognition of 200 Words, Int. J. Man-Machine Studies, 2, 223-234.