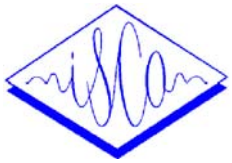

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MULTI-LINGUAL SPEECH RECOGNITION BASED ON DEMI-SYLLABLE SUBWORD UNITS

Tibor Fegyó and Péter Tatai
TSP Laboratory, Department of Telecommunications and Telematics,
Technical University of Budapest
Pázmány Péter Sétány 1/D, 1117 Budapest, Hungary
{fegyo, tatai}@bme-tel.ttt.bme.hu

ABSTRACT

Hungarian, unlike English, is an agglutinating language, so new, special methods are needed for speech recognition. The word dictionary could become very large due to its complex morphological system, so a suitable approach could be to use subword units as, eg., half (demi) syllables and language models. In this way most of the natural languages can be described, therefore this method can be applied in a multi lingual recognition system. We describe in the article a method to represent the demi-syllable segments efficiently. With this method the computation of the distortion measurement becomes fast. Our recognition results in speaker dependent case proved that the chosen method performs slightly better than DTW and requires sharply less computation time.

Keywords: cepstral trajectory, agglutinative languages, automatic segmentation, demi-syllables, multilingual

1. INTRODUCTION

In agglutinating languages, unlike in English, many prefixes and suffixes are used to modify the meaning of words. In particular, Hungarian is a strongly agglutinating language, more than a thousand different formative syllables and complex flexional endings [3] have to be considered even in a small or moderately sized dictionary of a recognition system. Clearly, in this case the number of full word models would become prohibitively large, resulting from the complex morphological system therefore subword models must be considered. In any case, if a large number of words have to be recognised, some kind of subword units are preferable to reduce the amount of computation needed in template matching.

At the TSP Lab a demi-syllable based speaker dependent recognition system is under development. As it is based on subword units, it can be readily adapted to various languages, either agglutinating or not, and allowing an easily extendable vocabulary [1].

Also, in a speaker dependent system it is essential to chose a pattern matching method which can be trained fast to every user with merely a few training samples. The standard DTW could be one possiblity, but it turned out to be quite slow and not sufficiently accurate [2]. On the other hand the statistical recognisers need too many training samples, so a new kind of template matching method is needed which will be presented in the following.

At first the automatic demi-syllable segmentation part of the recognition system is mentioned briefly [1]. Afterwards, the new template matching technique is presented which operates on demi-syllables. Describing the language model over these subword units, standard search algorithms are then used for word recognition. Finally, some experimental results are presented.

2. SEGMENTATION

A logical choice in many languages is the application of demi-syllables as subword units because their number is limited (typically less than a few thousands) and they seem also the best choice for automatic segmentation. The alternating statistical behaviour of the spectral parameters of human speech at the boundaries of demi-syllables can be efficiently used for automatic segmentation in a multi-pass recognition system [1]. The output of automatic segmentation is a sequence of so called demi-syllable like units, since their boundaries are not necessarily coincident with the linguistic units. In the template matching phase they can be handled as isolated units, therefore isolated word recognition methods can be applied in the recognition of continuous speech. Since this alternating phenomenon is present in most natural languages, the method can be applied in a multi-lingual recognition system.

According to preliminary experiments with a limited data corpus of Hungarian sentences (with a duration of 4...5 s each), the automatic segmentation in the first step has produced less than 10% segmentation errors (insertions or deletions) [1].

3. TEMPLATE MATCHING

Most automatic speech recognition systems use frame based modeling, where a frame is a 10-20 ms part of a speech sound, and these frames are treated independently. Another possibility is proposed in [4,5],

namely, the modeling several frames together, ie., a segment of a speech which can be a phoneme or any subword unit as well. Although this modeling is more realistic, its drawback is the high computational load. This is the practical reason of using the assumption of independent frames.

For the template matching phase we developed a preprocessor, which operates on demi-syllables instead of frames as classical front-ends. This new method is called Cepstral Trajectory Transformation (CTT) [2] because it transforms the time trajectories of the cepstral vectors of the segmented speech. The proposed model consists of two steps. In the first step the previously mentioned automatic subword (demi-syllable) segmentation is performed. Secondly a segment model is calculated from the frame based model, ie. cepstral vectors are calculated for each frame, and the trajectories of the cepstral vectors within a segment are determined.

With CTT the computation of distortion measures is particularly fast in comparison with dynamic time warping (DTW). Although in our preliminary recognition experiments the CTT is used with linear time alignment, the results are almost the same as those with DTW due to the short unit lengths. A further extension of the method will also be presented which applies nonlinear time alignment to achieve a slight improvement in the recognition rate.

Due to inevitable segmentation errors this method results in lower recognition rates than a statistical recogniser like a segmend based HMM, but the computation load is less, so it is more advantageous in real time systems.

4. CEPSTRAL TRAJECTORY TRANSFORMATION

The pattern matching and also the segmentation operates on cepstral vectors as in many other such systems. The cepstrum vectors are calculated for every 20 msec frame with 50% overlapping, and truncated to 20 components. One demi-syllable is described by 10-50 cepstrum vectors depending on the length of the segment. The time functions defined by the elements of the cepstral vectors are called trajectories. Our idea was to describe these trajectories with a limited number of parameters. The polynomial modeling and the discrete Fourier transformation were examined for realization, and finally the DFT was chosen [5]. For one demi-syllable the first three complex Fourier components gave sufficient accuracy. The components given by the DFT are called CTT parameters.

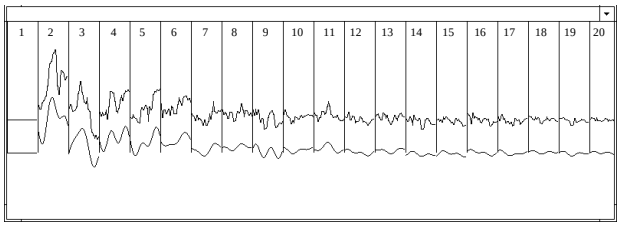

Figure 1. 20 trajectories and their smoothed versions of one demi-syllable

The result of the transformation is demonstrated in Figure 1. 20 cepstral parameters were used, so there are 20 trajectories next to each other (the upper curves). The approximation curves lay below the original trajectories, computed from the CTT parameters with inverse DFT. The approximation seems to be satisfactory, and can be considered as a kind of smoothing.

Fisher's discriminant

The generalized Fisher's discriminant [6] gives a solution for measuring the separability of feature elements. This discriminant was used to test the results of our method based on CTT parameters. The so-called *F* ratio is calculated as the variance of the means over all classes divided by the mean of the variances within classes. More exactly, the covariance for a single recognition class is defined by W_i where f_i are the vectors from this class and the feature centroid for $\overline{f_i}$.

$$W_i = E((f_i - \overline{f_i})(f_i - \overline{f_i})^T)$$

The pooled intraclass covariance matrix is *W*.

$$W = E(W_i)$$

Let $\overline{f}$ be the mean of f_i over all classes, hence the interclass covarince matrix is *B*.

$$B = E((\overline{f_i} - \overline{f})(\overline{f_i} - \overline{f})^T)$$

The generalized Fisher ratio can be calculated as

$$F = \frac{trace(B)}{trace(W)}$$

The *F* values are presented in Figure 2 for different cases defined in Section 6.

Distortion measurement

A distortion measure between two trajectories is needed for recognition. It can be calculated with the following formula, where *A* and *B* are the transformed CTT parameters, *O* is the number of the parameters (at present 3).

$$d = \frac{1}{O} \sum_{v=0}^{O-1} |A[v] - B[v]|^2$$

The distortion measure between two segments is the summed distortion of the trajectories. The closest reference to the segment is chosen as output. In the

basic method each training sample is stored for every demi-syllable, so if more than one training sample occur, all of them are stored as different references for the same demi-syllable.

Averaging the references

Generally for some of the demi-syllables there are more than one training sample even in a short training text. As an extension it is quite simple to combine them simply by averaging because the number of the CTT parameters are the same. In most cases the averaging results in higher recognition rates (see Table 1) because the different occurrences of the same demi-syllable are combined in one. For DTW it would be a more difficult problem, so in that case a common way is to store all of the training samples as references. The reference averaging applied is close to the mean value estimation in the HMM based systems, but the variance is not calculated because of the smaller number of training data.

In the case of CTT the basic advantage of the dynamic time warping is obviously lost if linear time alignment is used. However, a fast alignment could be used here as well with fixed (and only a few) number of parameters. The basic CTT provided slightly lower performance than DTW, but with reference averaging the results were similar despite of the linear alignment. This fact may be due to the smoothing effect of the CTT method and the averaging. Also, the demi-syllables are too short to exhibit large dynamic changes along the time scale. Nevertheless, some experiments were performed to exploit the possible improvements by using non-linear alignment.

Dynamic extension of the CTT

In the confusion matrix it was found that in most cases the vowel part of the demi-syllables was recognised correctly, but the consonant part was not. This phenomenon is due to the characteristics of the pattern matching which is dominated by the spectrally constant part of the speech, since the non-constant part is considerably shorter. Taking only the vowel part of the demi-syllables higher than 80% recognition rates are obtained (see Table 2).

To reduce the effects above and to provide dynamic time handling an extension of the CTT was also investigated. A non-linear function on the time scale of the cepstrum vectors was applied before the CTT. This transformation is a function of the grade of the change of the parameters. In our case this parameter (ΔP) is the average of the change of the cepstrum components.

$$\Delta T_i = \{C_0 + C_1 * f(\Delta P_i)\} T_0$$

This function emphasise the changing part of the signal, and the constant parts get less weight. The amount of emphasis is a function of the parameters C_0 and C_1. After finding $C_1 / C_0 \cong 10$ as optimum, this modification provided slightly better recognition results and better discrimination abilities than the simple linear time warping. The computation was still significantly faster than that of the DTW. Our time alignment operated best when the quickly changing part, ie., the consonant part of the demi-syllables was emphasised,, so the consonant recognition rate increases assuming that the segmentation is correct.

With this kind of emphasis, however, the method became more sensitive to the segmentation errors. If a segmentation point is not exact enough, the procedure emphasises this error by giving more weight to this changing part.

The mentioned three CTT algorithms (basic, +averaging, +non-linear alignment) have been implemented and tested with the same words as a DTW. The results are promising. The basic CTT provided slightly worse performance than the DTW, with significantly shorter processing time. The CTT with reference averaging and nonlinear time alignment performed even better than the DTW. The results are demonstrated in Section 6.

5. WORD RECOGNITION

For languages with an extensive inflectional and derivational morphology like Hungarian, the language model may need to be defined at the morpheme level to allow decreasing the size of the dictionary. Language model described by regular expressions or an N-gram model, is an ideal tool for the recognition of agglutinative languages [3].

As a preliminary experiment the HTK toolkit [7] was used to recognise words. Each subword unit was represented by a special HMM. Having the language model: word and morpheme dictionaries, morphological rules and standard HMM search techniques were applied at the word level matching. However, the segmentation and the first pass recognition phase could work efficiently with one or just a few training samples therefore the fast training for speaker dependent recognition has been retained.

6. EXPERIMENTAL RESULTS

All of the tests were performed on the same Hungarian data set of 32 different demi-syllables. The data set included 200 words and 720 demi-syllables. 10% of them were used for training, and the others for testing.

Measure of separability

At first the measure of separability [6] was calculated by applying the generalized Fisher's ratio for different subsets, both on the CTT parameters and on the preemphasised CTT parameters.

The results are presented in Figure 2; 'total' means the

whole data set, 'vowel', only the vowels from the whole set. These vowels are *a* [a], *á* [a:] and *o* [o]. Then the data were separated according to the vowels, and the *F* value was calculated for each group. For the CTT 1.54 was obtained and for the preemphasized version a slight improvement to 1.56 was achieved. These numbers are over the breakeven point [6], which is 1.0 for this measure, so the feature set is acceptable. For vowels higher values are obtained as it was expected. More interesting question was to verify this ratio for similar demi-syllables, therefore group *a, á* and *o* were examined separately. For the vowel *a* and *o* the results were promising as well, but at group *á* it was quite low, which requires further investigation of the underlying reasons.

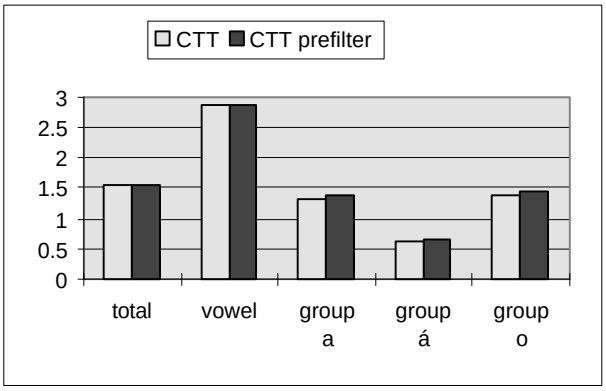

Figure 2. The *F* values for different data sets.

Demi-syllable and word recognition

The results of the detailed recognition experiments are presented in Table 1. At first the basic CTT was tested. Every training sample was stored as reference. This experiment proved that a reference from a single utterance is not always satisfactory, recognition rates of the subword units were in the range between 5% and 89%. With averaging two or three samples the improvements achieved fell between 20-200% (depending on the demi-syllables). Note that this method does not require a large number of training samples as a statistical method, and two-three samples already result in a significantly better performance than one. If the number of training samples are increased further, this new method becomes very robust and can be applied efficiently even in a speaker independent recognition system.

At the word level recognition the HTK toolkit was used [7]. The HMM models had only one stage, the mean vectors were the same as those at the previous CTT experiment, and the variances were unity. The 100 words tested were constructed from the 32 demi-syllables. The word recognition rates are 4% higher than for the individual demi-syllables. The relatively small increase is probably due to the particular vocabulary which contains many similar words.

DTW	Base CTT	Averaged CTT	Averaged CTT with preemphasis	HTK word modells
half syllable recogniton				word rec.
52,2%	42,7%	54,5%	56,3 %	60,8 %

Table 1. Averaged recognition rates

Vowel recognition rates were calculated also from the confusion matrix (Table 2). In most cases the vowel part was recognised correctly, only the consonant part was missed. So the vowel recognition rates are quite high. As expected, the non-linear CTT provided about the same vowel recognition rates as the linear one.

The system was trained quickly and automatically for a single user, but in this process incorrect training samples occur as well, which decrease the recognition rate. In the future an interactive training is planned or precollected references should be used with user adaptivity.

DTW	Base CTT	Averaged CTT	Averaged CTT with preemphasis
84,80 %	79,64 %	87,87 %	88,28 %

Table 2. Vowel recognition rates calculated from the confusion matrix.

7. SUMMARY

This paper proposed a method for pattern matching, called Cepstral Trajectory Transformation which is fast enough for real-time applications. The generalized Fisher's discriminant was calculated on the feature set provig that the discrimination ability is acceptable. In the first tests a basic CTT is used with linear time alignment, which provided slightly worse results than a DTW. Therefore a reference averaging extension is introduced, which increases the recognition rate in most cases. Then the performance became even better than that of the DTW. To emphasise the consonant part of the demi-syllables a non-linear extension is also tested which improved the recognition rate further.

8. REFERENCES

[1] Bán, L. and P. Tatai (1997), Segmentation for an Open Vocabulary Recognition System. *Proceedings of The first European Conference on Signal Analysis and Prediction*, Prague, pp. 303 - 306.

[2] Fegyó, T. and P. Tatai (1998), Cepstral Trajectory Transformation for Subword Recognition. *Proceedings of the Workshop on Text, Speech and Dialogue*, Brno, pp.189 - 194.

[3] Tompa, J (editor) (1961), System of the Hungarian language. Akadémiai Kiadó, Budapest (in Hungarian)

[4] Szarvas, M. and S. Matsunaga, (1998) Acoustic

Observation Context Modeling in Segment Based Recognition. *Proceedings of the ICSLP-98*, vol. VII., pp.2967 - 2970.

[5] Gish, H.and K. Ng (1993), A segmental speech model with applications to word spotting. *Proceedings of the ICASSP'93*, pp. II-447 - II-450.

[6] Parsons, T (1986), Voice and Speech Processing. McGraw Hill, pp. 170 - 194.

[7] Young, J.S. (1990), HTK: Hidden Markov Model Toolkit v1.2. Reference Manual, Cambridge University.