EXPERIMENTS ON THE USE OF DEMISYLLABLES FOR AUTOMATIC SPEECH RECOGNITION

G. Ruske

Lehrstuhl f. Datenverarbeitung, Techn. University of Munich, Franz-Joseph-Str. 38, D-8000 Muenchen 40, Fed. Rep. of Germany

Abstract: The paper describes methods for an explicit segmentation of the speech signal into demisyllable segments by evaluating the output of a loudness model. Syllable nuclei are indicated by maxima of a smoothed loudness function. Consonant clusters and vowels are introduced as decision units in order to reduce the inventory of classes. Two methods for classification of consonant clusters are compared: template matching and a feature extraction approach based on acoustic cues. Sentence recognition operates on phonetic word models adapted to the demisyllable structure.

1. INTRODUCTION

An important question in automatic speech recognition is the choice of basic units which have to be processed basically by the system. A segmentation procedure tries to divide the speech signal into individual parts (segments) in such a way that they can be processed as independently as possible. The segmentation can be performed implicitly when classification of the segments and determination of the segment boundaries are carried out in common. However, this usually requires an enormous expenditure of computing power. On the other hand, segmentation can be carried out explicitly by placing definite segment boundaries in the speech signal; classification now only has to treat the fixed segments. In this case, however, the system must be prepared for the fact that the segmentation step may cause errors, too. The subsequent stages of the system have to be able to correct these segmentation errors (see Sect. 5).

The speech recognition system described in this paper starts from an explicit segmentation into demisylables. These processing units have the advantage that the main coarticulation effects are contained within the segments. The number of classes can be drastically reduced when consonant clusters and vowels are used as decision units for classification.

Evaluation of the syllable structure in the speech signal is facilitated by using a loudness model of hearing /1/ for preprocessing. This model consists of a critical-band-rate filter bank with 24 band-pass filters; 22 channels are used in the system (50 Hz - 8.5 kHz). All channels are processed by a loudness model which simulates the masking effects in hearing. The outputs of the model are sampled every 10 ms; the 22 components constitute a so-called loudness spectrum, see fig. 1a. The total loudness N(t) is calculated as the sum of all 22 components; additionally a weighted sum of these components gives the so-called modified loudness $N_m(t)$ which is very useful for syllabic segmentation. Fig.1b displays the block diagram for the calculation of these functions.

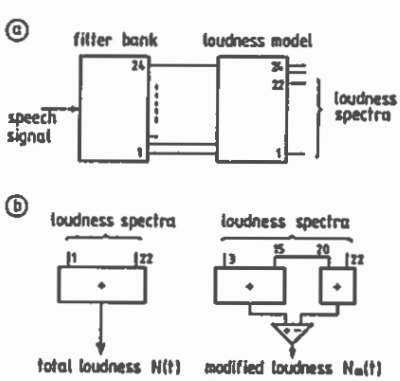

Fig. 1. a) Block diagram of preprocessing; b) calculation of N(t) and $N_m(t)$.

2. DEMISYLLABLE SEGMENTATION

The modified loudness $N_m(t)$ evaluates the frequency range which is dominated by the vowels. Therefore this function is especially suited to indicate the syllable nuclei (vowels and diphthongs). When this function is smoothed according to the average syllable rhythm in the speech signal, the local maxima of this function indicate the positions of the syllable nuclei. For this purpose a special smoothing filter (digital low-pass filter) has been applied having a Gauss-like impulse response h(t), see fig. 2; in the digital calculation this function corresponds to h(n) with $n = n \cdot \Delta t$ (10ms). This smoothing filter has been realized on the basis of an elementary filter with a rectangular impulse response; the output sample y(i) is calculated from the input signal x(n) as:

$$y(i) = 1/3\ (x(i-1) + x(i) + x(i+1)) .$$

When this filter is placed k-times in series, the impulse responses of fig. 2 result. The repeating factor k now determines the time constant T of the filter, see fig. 2. This smoothing filter is applied to $N_m(t)$. The time constant T has been optimized using test material consisting of 23 sentences spoken six times; the speech material contained 2566 syllables altogether /2/. It is important to adjust the time constant T to the speaking rate: for a short time constant T many surplus syllable nuclei are marked (insertions), for long time constants T many nuclei are smoothed out resulting in omissions. Both effects contribute to the total segmentation error rate as depicted in fig. 3. It can be seen from the figure that an optimal value for T was reached for k=7 corresponding to a time constant T=55.7 ms (this is equivalent to a cut-off frequency of the filter f = 9 Hz). The minmum error rate was 3.66% (from 2566 syllables /2/). It has to be borne in mind that here only the maxima of the optimally smoothed function $N_m(t)$ were evaluated. A further reduction in the segmentation error rate is achieved by evaluating the spectral information at the positions of the maxima indicated by $N_m(t)$ /3,4/. As an extreme solution, a complete vowel classifier can be applied at each time instant in order to estimate the syllable nuclei /2/. In the realized recognition system a combination of both methods was implemented which has a segmentation error rate of about 4-8% in practical applications with continuous speech.

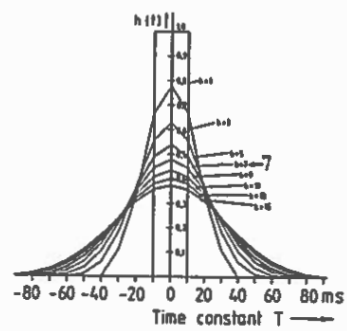

Fig. 2. Impulse response h(t) (from /2/).

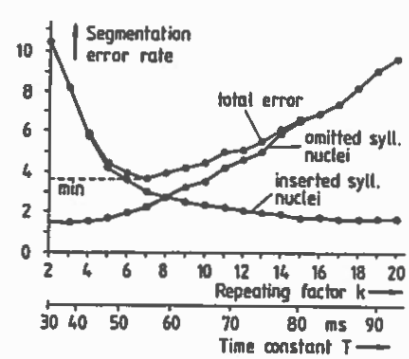

Fig. 3. Segmentation error rate for syllable nuclei as a function of T (from /2/).

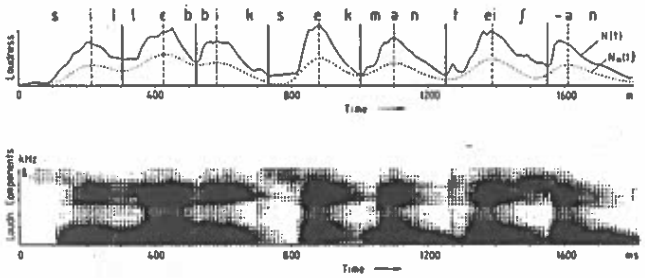

Fig. 4. Demisyllable segmentation of the utterance "syllabic segmentation".

Syllable boundaries are placed at local minima in the loudness N(t) between two consecutive syllable nuclei. When more than one minima are present, the lowest minimum is chosen /3/. This method yields in most cases a suitable boundary. The demisyllable segment now spans the range from the syllable boundary to the syllable nucleus, see fig. 4.

4. CLASSIFICATION OF CONSONANT CLUSTERS

Each demisyllable segment contains a consonant cluster and a part of the vowel from the syllable nucleus. The number of different units can be drastically reduced when consonant clusters and vowels are introduced as decision units for the classification. In the German language we only have to discriminate about:

- 50 initial consonant clusters,
- 20 vowels (inclusive diphthongs),
- 160 final consonant clusters.

That means that the demisyllable is seen as a segmentation and processing unit but not as a decision unit for recognition. In this way the huge inventory of different demisyllables can be avoided while preserving the advantages of demisyllable segmentation.

4.1 Classification by template matching

A first approach to recognition of the consonant clusters consists in using complete spectral-temporal templates of all consonant clusters. For this purpose a special time normalization procedure was developed called "dynamic interpolation". Details of this procedure have been described in /3,4/. After normalization of the demisyllable segment, a city-block metric can be applied for the calculation of similarity.

Experiments have been carried out with a test corpus of 368 initial and 384 final demisyllables which were automatically segmented in German words spoken by one male speaker /5/. This material contained 45 initial consonant clusters and 48 important final consonant clusters. The average recognition score using the template matching method amounted to 66% for initial and 75% for final consonants. These results can be seen as good as those typically obtained in automatic consonant recognition. Vowel recognition will not be discussed here.

4.2 Classification by feature extraction

A second approach starts from a description of those acoustic events within a demisyllable that are relevant for phonetic decoding. For this purpose the following features or "cues" were measured: formants, formant transitions, formant-like links for nasals and liquids, turbulences (or bursts), pauses, and voice-bar within pauses or turbulences. These cues are characterized by spectral and temporal measurements. Since the number and order of consonants is restricted in syllable-initial and final position, initial consonant clusters could be completely described by 24 feature components and final consonant clusters by 31 components /5/.

The feature extraction methods are based on the evaluation of energy in several spectral bands and are described in /6/. The context dependencies are taken into account by collating all feature components derived from a demisyllable segment into a common feature vector. For comparison, this method was applied to the same speech material (see Sect. 4.1).

From the recognized consonant clusters the recognition scores of the single consonants were computed. The recognition scores were 4 and 7% lower as compared with the template matching approach /5/. However, it has to be borne in mind that the feature vectors consisted only of 24 or 31 components whereas the templates needed several hundred components for their representation. Thus the feature approach can indeed be seen as a suitable basis for the acoustic-phonetic analysis of demisyllables.

5. RECOGNITION OF SENTENCES

Demisyllable segmentation and recognition has been incorporated in a system which processes spoken sentences as a chain of connected words. This system is completely described in /7/ and will be summarized here only very briefly.

Each word of the vocabulary is represented by a phonetic word model containing the variations in pronunciation as well as possible segmentation errors. The models are constructed in such a way that they can be processed very efficiently by use of Dynamic Programming (DP) methods.

Sentence recognition is based on a 1-stage DP algorithm which determines the best match between a series of word models and the phonetic symbols (consonant clusters and vowels) provided by the classification stage. The word models and the DP transition rules take particular account of the syllabic structure of the utterance.

First experiments with a 75 word vocabulary resulted in recognition scores of 85% correct words in continuous speech without utilizing any grammatical or semantic information. These encouraging results demonstrate the efficient use of syllabic units in all stages of a speech recognition system.

References:

/1/ ZWICKER, E., Peripheral preprocessing in hearing and psychoacoustics as guidelines for speech recognition. Symp. Montreal 1986, Proc. of this conf.

/2/ GEYWITZ, H.-J., Automatische Erkennung fließender Sprache mit silbenorientierten Einheiten. Doct. Thesis, Techn. Universität München, 1984.

/3/ SCHOTOLA, T., On the use of demisyllables in automatic speech recognition. Speech Communication 3, Elsevier Science Publ., 1984, 63-87.

/4/ RUSKE, G., Demisyllables as processing units for automatic speech recognition and lexical access. In: "New Systems and Architectures for Automatic Speech Recognition and Synthesis" (R. DeMori and C.Y. Suen, eds.), Springer-Verlag, 1985, 593-611.

/5/ RUSKE, G., On the usage of demisyllables in automatic speech recognition. In: SIGNAL PROCESSING II: Theories and Applications, (H.W.Schüßler, Ed.), Elsvier Science Publ. (North-Holland), 1983, 419-422.

/6/ RUSKE, G., Automatic recognition of syllabic speech segments using spectral and temporal features. IEEE ICASSP, Paris, 1982, 550-553.

/7/ RUSKE, G. and WEIGEL, W., Automatic recognition of spoken sentences using a demisyllable-based dynamic programming algorithm. 12th ICA, Toronto, July 1986, in print.