Syllable-based Analysis of Spectral and Temporal Features for Automatic Speech Recognition

G. Ruske
Munich, FRG

1. Introduction

In systems for automatic recognition of continuous speech an initial segmentation is required in order to obtain appropriate segments which can be used as basic units for the subsequent analysis and classification procedures. The syllable structure of the speech signal provides segments which include the essential coarticulation effects. A substantial reduction in the number of different syllables is achieved by a further subdivision of each syllable segment into two parts: an initial demisyllable ranging from the starting point of the syllable to the syllable nucleus, and a final demisyllable ranging from the nucleus to the end point of this syllable. In the German language we only have to discriminate about 50 initial consonant clusters, 19 vowels (short and long vowels, and 3 diphthongs). and there is a maximum of about 160 final consonant clusters. The efficiency of syllabic segmentation has been demonstrated in previous experiments (Ruske and Schotola, 1981).

2. Demisyllable segmentation

A syllable is defined here as an 'acoustic syllable'. According to this definition the localization of syllable nuclei as well as the determination of a suitable syllable boundary between two consecutive syllable nuclei can be based on an evaluation of the loudness contour and on spectral information from the speech signal. In our experiments we examined several methods for syllabic segmentation of spoken utterances:

a. The maxima of a smoothed loudness function have proved to be suitable candidates for syllable nuclei. An additional vowel/non-vowel classification of the spectrum at the maximum rejects maxima produced by high energy consonants.
b. A demisyllable boundary is indicated by a loudness minimum.
c. A crude vowel classification is performed at intervals of 10 ms using several vowel spectra as prototypes. High estimates of this classifier indicate syllable nuclei.
d. A set of demisyllable templates is applied consisting of spectral-temporal patterns and which are representative for all possible demisyllable shapes. The comparison is performed by dynamic programming methods (2-level DP matching). This method provides syllable nuclei as well as the syllable boundaries.

Methods a. and b. were used for the following experiments, since they yield reliable segmentation results and in addition save computation time.

3. Recognition of demisyllables

As a first approach, recognition of German demisyllables was carried out using spectral-temporal templates of complete consonant clusters and spectral templates of vowels. Time normalization was performed by a so-called 'dynamic interpolation' procedure. After normalization a city-block metric was applied for the calculation of similarity. However, template matching needs a lot of storage and computation time since an unknown consonant cluster has to be compared with all reference templates regardless of their phonetic structure.

Since knowledge about the gross phonetic structure could considerably reduce the number of templates to be actually compared a second method was developed starting from a description of the relevant acoustic events within each segment by evaluating spectral and temporal features or 'cues' which can be objectively measured in the signal. These features have been defined in some analogy to the classical perception experiments with synthesized speech sounds which have been reported in the literature (Delattre, 1968). The cues describe: the 'loci' of the first 3 formants of the syllable vowel, the formant transitions, formant-like 'links' for nasals and liquids, duration and spectral distribution of bursts and turbulences, pauses, and voicing during pauses or turbulences.

A main problem is the depency of most of the features on phonetic context. In the present paper the context dependencies are taken into consideration by collating the results of feature extraction within each demisyllable segment. This enables the contextual dependencies between the acoustic features to be determined statistically from representative speech material. The feature vector for an initial consonant cluster as well as for a final demisyllable have a fixed number of components. In syllable-initial position 1 nasal, liquid or glide, and up to 2 fricatives or plosives are possible; in syllable-final position the maximum number of plosives or fricatives can be limited to 3. Therefore, initial consonant clusters are completely described by 24 feature components and final consonant clusters by 31 components, see Table Ia and Table Ib.

4. Feature extraction method

Feature extraction starts from a spectral representation calculated by linear prediction analysis. From the LPC-coefficients power spectra are computed with a frequency resolution of about 78 Hz. Depending on a voicing parameter each spectral frame is labelled voiced, unvoiced, or silent, respectively.

Table 1a. Feature vector for initial consonant clusters

Component number	Feature	Dimension
1	First turbulence (or burst)	Yes/No
2	Center of gravity	Hz
3	Lower cut-off frequency	Hz
4	Upper cut-off frequency	Hz
5	Duration	ms
6	First pause	Yes/No
7	Duration of pause	ms
8	Second turbulence (or burst)	Yes/No
9	Center of gravity	Hz
10	Lower cut-off frequency	Hz
11	Upper cut-off frequency	Hz
12	Duration	ms
13	Voice-bar	Yes/No
14	Duration of voice-bar	ms
15	Nasal or liquid links	Yes/No
16	Low link	Hz
17	High link	Hz
18	Transition of links	Hz/ms
19	Transition of the formant F_1	Hz/ms
20	Transition of the formant F_2	Hz/ms
21	Transition of the formant F_3	Hz/ms
22	Locus of the formant F_1	Hz
23	Locus of the formant F_2	Hz
24	Locus of the formant F_3	Hz

Finally, formant tracking is performed within voiced parts. Acoustic parameters are then derived from the energy in selected frequency bands which allow a gross characterization of the spectral shape; the feature extraction procedures are in some accordance with the methods proposed by Weinstein et al. (1975). Based on these parameters, a set of rules has been established in order to detect voiced and unvoiced turbulences and bursts, pauses, and liquid and nasal links. After detection the individual features are characterized by gross measurements of their spectral and temporal distribution (e.g. center of gravity, upper and lower cut-off frequency for turbulences, and spectral peaks for links) and used as components of the common demisyllable feature vector.

5. Classification procedure

Classification of the feature vectors is based on Euclidean distance measurements within the feature space. All components are normalized to equal variance. However, when comparing the unknown vector with a reference vector, the single components are handled in different ways. For all binary components which indicate the presence or absence of a single feature, the

Table 1b. Feature vector for final consonant clusters

Component number	Feature	Dimension
1	First turbulence (or burst)	Yes/No
2	Center of gravity	Hz
3	Lower cut-off frequency	Hz
4	Upper cut-off frequency	Hz
5	Duration	ms
6	First pause	Yes/No
7	Duration of pause	ms
8	Second turbulence (or burst)	Yes/No
9	Center of gravity	Hz
10	Lower cut-off frequency	Hz
11	Upper cut-off frequency	Hz
12	Duration	ms
13	Second pause	Yes/No
14	Duration of pause	ms
15	Third turbulence (or burst)	Yes/No
16	Center of gravity	Hz
17	Lower cut-off frequency	Hz
18	Upper cut-off frequency	Hz
19	Duration	ms
20	Third pause	Yes/No
21	Duration of pause	ms
22	Nasal or liquid links	Yes/No
23	Low link	Hz
24	High link	Hz
25	Transition of links	Hz/ms
26	Transition of the formant F_1	Hz/ms
27	Transition of the formant F_2	Hz/ms
28	Transition of the formant F_3	Hz/ms
29	Locus of the formant F_1	Hz
30	Locus of the formant F_2	Hz
31	Locus of the formant F_3	Hz

distance is calculated in each case. The distances of the remaining components describing the temporal or spectral characteristics of a certain feature are only taken into account in those cases where the corresponding acoustic feature is present in both vectors. Finally the accumulated distance is divided by the number of all feature components evaluated during the distance measurement. This results in the calculation of an 'average normalized distance' which allows the comparison of all feature vectors even if they differ quite considerably as to their current composition.

6. Recognition experiments

The speech test material consisted of several hundred initial and final demisyllables which were automatically extracted from German words spoken by

one male speaker. The set of demisyllables contained 45 initial consonant clusters and 48 important final consonant clusters, all combined with 8 vowels. Syllabic segmentation, feature extraction and classification was applied to this material. As expected some difficulties arise in the detection of the glide /r/ and the liquid /l/. The consonant /r/ often cannot be discriminated from the vowel, whereas in the case of /l/ often parts of the vowel /o/ or /u/ were indicated erroneously as liquid links as e.g. in /ʃu:/.

Some of the confusions observed in the feature classification experiments can be explained by inspection of the mean values and standard deviations. The data display the typical order of the F_2-loci for the plosives: low for the labial /p/, mid for the dental /t/ and high for the velar /k/. However, the standard deviations are rather large so that the corresponding confusions are to be expected. A special problem is the discrimination between /m/ and /n/ which were often confused. On the other hand, the calculation of the gradient of formant-to-link transitions enabled nasals to be discriminated from liquids. From the recognized consonant clusters the recognition scores of the single consonants were computed; the average recognition rate for the single consonants was about 62% for initial and 68% for final consonants.

For comparison, the same speech material was processed by *template matching* methods using complete spectral-temporal templates for each consonant cluster. Here, the average recognition score was about 4-7% better and on average amounted 66% for initial and 75% for final consonants. Again confusions occurred between the unvoiced and voiced plosives, and between the nasals and /l/ and /v/. The fricatives received the best recognition scores. It is worth noticing that, roughly speaking, the distribution of confusions obtained by template matching is very similar to or even identical to that obtained by feature extraction. While the feature extraction approach could not yet reach the recognition accuracy of template matching, it has to be borne in mind that the feature vector for a consonant cluster has only 24 or 31 components whereas a corresponding template constructed from a series of consecutive spectra needs on average more than 500 components; this results in about 20 times more storage and computation time. Thus the features components can be seen as an efficient representation of the units. In both experiments the recognition scores were not very high; they have to be seen as pilot experiments. But the main goal of this investigation was only to compare the efficiency of the two methods. Our own previous investigations (with template matching) showed that an 85-90% consonant recognition score can be reached with a large training set. This encourages us to believe that the recognition scores of the feature approach can be considerably improved by further optimizing the feature extraction procedures.

References

Delattre, P.C. (1968). From acoustic cues to distinctive features. *Phonetica* **18**, 198-230.

Ruske, G. and Schotola, T. (1981). The efficiency of demisyllable segmentation in the recognition of spoken words. *IEEE ICASSP* 1981, 971-974.

Weinstein, C.J., et al. (1975). A system for acoustic-phonetic analysis of continuous speech. *IEEE Trans. Vol. ASSP-23*, 54-67.