DYNAMIC DETERMINATION OF ACOUSTIC VOCAL CONTRAST

FLOWER J. KOPPANSA - VAN BEINUM

ROB P. DE SAINT AILHARE

Institute of Phonetic Sciences, University of Amsterdam.

ABSTRACT

A problem in automatic speech recognition as well as in speech synthesis-rule is how to cope with the phenomena of vowel reduction: vowels in connected speech rarely reach their target position (the intended vowel) as defined in isolated- and word-integrated vowels. In this paper, a novel automatic dynamic procedure for determining vowel contrast (ASD-

contrast), a dynamically adjustable global measure for acoustic system contrast (ASC). This global ASC-measure, combined with local parameter values in the dynamic vowel analysis, may provide in due time various applications with respect to the description and use of vowel reduction aspects. On the basis of connected speech material (read tests and free conversation) of both Dutch and one Japanese speaker, the present results are compared with similar data earlier derived by hand segmentation and average vowel formant data per vowel segment.

INTRODUCTION

The great variability in the realization of vowel sounds, when produced by the same speaker, is a role in speech synthesis as well as in speech recognition. Vowels rarely reach their target position (the intended vowels) as defined in isolated- and word-integrated vowels. In speech synthesis, we may need a model to describe this variability in order to increase intelligibility as well as naturalness. In automatic speech recognition, vowel reduction is an annoying phenomenon.

In connected speech, the degree of acoustic contrast between vowels in a speaker's vowel system is dependent on a number of factors, e.g., speaker, articulatory region, and local context, quite a lot of research has been done with respect to the description of vowel contrast in various speech situations. However, the relations between those factors are more specifically their hierarchical structure has been studied only fragmentarily. All words (1963) for instance postulates that duration in the context of connected speech, whereas De Laat (1963) claims stress and speech rate are not determinants with duration as a product of stress and tempo and therefore a secondary determinant. Boy (1977) and Den Oe (1985) both show that an increase of speech rate not necessarily affects the present vowel sounds. Furthermore Kopperna and Van Beinum (1985) indicate a different relation between stress and vowel duration for read tests as compared to tests with a free choice of words (retold story or free conversation).

As for perceptual studies in stress (e.g., Van Eeke and Dedeme, 1974; Rietveld and Kopperna and Van Beinum, 1983) to appear) the relation between loudness, intonation, speech rate, and vowel contrast reduction turns out to be quite complex.

In order to reach a better understanding of the relation between the hierarchical structure of the great variability of vowels, it is deemed necessary in our approach of both global factors and local factors. The former factors in connected speech is divided into a speaker's vowel system and connected speech.

We therefore started a project in order to develop and apply strategies to make optimal use of acoustic, acoustic-phonetic information if possible also linguistic information with the global aspects with respect to the local analysis of connected speech.

This will be done by means of a dynamic acoustico-phonetic method for dynamic vowel analysis and cumulative data processing.

a) Any speech fragment of any speaker may be submitted to a dynamic acoustico-phonetic analysis to provide information on global aspects such as articulation, vowel space, overall speech rate, degree of vowel contrast, and the acoustical pitch values in the dynamic vowel analysis. It will also show how the moment when the global measure for acoustic system contrast (ASC-stabilizes) indicates the duration of the speech sample needed for defining this value (other measures, like articulation, are dynamically adjusting it, if use is made of a moving window).

b) Subsequently local acoustic vowel contrast or degree of reduction and variability will be derived from dynamic-acoustic-phonetic parameters as fundamental frequency, formant frequencies, bandpass filters, values, vowel duration, amplitudes.

c) Finally the results of a) and b) will be used in various applications, as for instance labelling of segments as specific vowel phonemes, merely by using the local analysis and cumulative data processing combined with global contrast measures and general information on the present vowel sound, and defining the hierarchical structure of factors influencing the variability in vowel phonemes.

This paper reports on our first steps within this project, i.e., the development of a method for the dynamic determination of the global measure for acoustic vowel contrast and its application to two quite different languages, Dutch and Japanese. In three main questions have to be answered:

1) What is the global measure for connected speech?

2) How does the global measure compare with the traditional static measure?

3) Is the dynamic (semi-instantaneous) procedure applicable to two phonetically quite different languages.

SECTION OF A DYNAMIC ANALYSIS AND DATA PROCESSING METHOD

As the aim of the present subproject was to develop a (semi-)automatic methodology of processing data, two parallel methods had to be developed: a) the traditional method making use of manual segmentation of vowels in the digitized speech signal fragment by means of a speech editor, followed by a dynamic acoustic-phonetic analysis, and b) a (semi-)automatic method by referring to a dynamic acoustic analysis, first to the entire speech sequence, followed by an automatic vowel selection. Both methods are followed by a data processing step based on formant frequencies or basis functions or band-pass values, which calculates in both cases the acoustic system contrast measure ASC (Kopperna and Van Beinum, 1980). This ASC measure is defined in the form of the temporal-vowel system, based on frequencies of the first (F1) and second formant (F2), transformed to 100 log Hz, using the formula:

\[ ASC = \frac{1}{1 + \frac{V_{1}}{V_{2}}} \]

where V1 and V2 are the 2-dimensional vector of the vowel j in the F1-F2 plane, C1 is the 1-dimensional vector of the centroid C, vj is the number of vowels in the vowel system.

d) A second step of the system contrast measure of ASC of any specific speaker is defined now by the total variance of all formants in 10 ms formant frames, just as they occur in the speech fragment. The moment at which this ASC stabilizes actually defines the length of the speech fragment.

It is claimed that in so-called syllable-based languages, like e.g., Spanish, Italian, and Japanese, the degree of spectral reduction is much less, if present at all, than in so-called stress-based languages like e.g., English, Russian, and Dutch. In previous work, however, we noted for Dutch and for Japanese there is a degree of vowel reduction expressed in comparable values of acoustic system contrast (Kopperna and Van Beinum, 1982/83). Therefore we decided to test our dynamic analysis methodology on Dutch as well as on Japanese speech material.

The Dutch vowel system consists of twelve short or long monophthongal vowels and three diphthongs. All vowels and diphthongs may occur in stressed or in unstressed position. Furthermore 3 of all vowel phonemes in Dutch consists of actual sounds apart from reduced vowel sounds. The schwa phoneme occurs only in unstressed position. The diphthongs are longest in duration, then there are four long monophthongs, and the remaining vowels including the schwa sounds are the shortest. The Dutch short vowel sounds are not completely realized than the long vowels (for more details see Kopperna and Van Beinum, 1980).

The Japanese vowel system is rather simple one consisting of only five vowels, according to Takebayashi (1975) the vowels /i/ and /u/ are often devoiced when they occur between voiceless consonants and in word-final position. Stress does not seem to play any phonological role in Japanese and it is claimed that all vowels always pronounced without narrow phonological contrast. However, the pronunciation of the latter class is proved by De Graaf and Kopperna and Van Beinum (1982/83) who demonstrated a similar degree of reduction in connected speech for a number of languages and a number of dialects.

As for Dutch we used recorded speech material of the same trained male speaker and the same native speaker of Japanese (Kopperna and Van Beinum, 1980). This provided us with the possibility of a comparison of the results of the automatic procedures with previous results. Nevertheless an important difference remains in the present speech material. The recorded speech material was manually stylized as being vowel-like in the chosen speech fragment, and the automatic vowel reduction was based on the way they occurred. Moreover measurements were carried out only dynamically with time intervals of 10 ms which means that frequency of occurrence of all vowels in normal running speech of a speaker is not preserved, and that the duration of each occurring vowel was estimated on the basis of the overall acoustic system contrast. In the former study, all vowels in every fragment were used and were measured only as a point sound which means that it is no longer necessary to 'label' the vowel when the specific speaker is defined by the specific speaker he is not specific to the specific speaker but only the specific speaker he is not specific to the specific speaker but only the specific speaker.
needed for the determination of the ASC for that specific speaker in that specific speech situation. As to how far depth of fragment dependence of speaker, on speech acoustics, and on language is one of the research questions of the proposal.

As for the Dutch speaker we made use of two situations: a free conversational (30 sec fragment) and read text (a 10 sec fragment). The speech frequencies were obtained from existing recordings, as shown which provides us with the possibility to compare the static and dynamic analysis method using exactly the same fragment (not the same recording, not exactly the same fragment). Our decision to confine ourselves to a 30 sec fragment is based on literature indicating that variables concerning the distribution of speech energy stabilize within a period of time (Li, Hughes, and House, 1969; Zaborin and Rothenberg, 1963). Our choice of only a 10 sec fragment of read text is defined by the results obtained from the free conversational fragment and the need of confining the material.

As for Japanese the speech material of one male speaker (free conversational and read text) has been recorded in Japan recently. Since this speaker was not involved in the earlier studies on the Japanese vowel system, comparison of the earlier analysis results with the results of the dynamic analysis would not make much sense. Therefore in order to answer our questions and at the same time to limit the analysis material we confined ourselves to Dutch conversation (30 sec, 20 male speakers, 10 per sex, manual and automatic), Japanese conversation (10 sec, 20 male speakers) and Japanese read text (10 sec, manual and automatic).

**MEASUREMENTS**

By means of a speech editing program (Butting, 1968) all vowel recordings were isolated in such a way that it was clear that the place of the voice production and the vocal tract were clearly visible in the oscillogram for the first and last 15% of the recording. It was considered to be the case where the points in three fundamental frequencies were visible in the oscillogram for the first third and the last third of the recording. For the Japanese the same was considered. From the visual appearance of the oscillograms, we identified those cases where the specific formant pattern disappeared. In many cases the visual appearance of the formant pattern indicated that the successive samples were segmented that did not for this reason any truly visually observable consonant form. Once the total formant pattern had been identified, all vowel recordings of course were known as well. From the LVA recordings 121 vowels could be selected with an average duration of 1000 ms each. The fragment of read text 57 vowels were segmented according to the same criteria. The manual segmentation procedure has been analyzed dynamically in 10 ms steps. A set of limited conditions as a spectral analysis program called Q (Street, 1965) was used as the standard.

Apart from a number of other procedures, not relevant for this study, the program QQ provides us with: fundamental frequency measurements, calculation of normalized acoustic-algorithm (Butlins, 1967),

- fundamental frequency: a band filter analysis base on the FFT amplitude spectrum in a critical band with filter specifications given by Sekey et al.

The result values are stored in analysis file consisting of records, each of which contains the analysis result of one 10 ms voice frame, in this way all kind of selection can be made.

The measurement of a development of an automatic procedure of data processing, one of the main problems was the automatic segmentation from the speech fragment (cf. Kanya, and Nakita, 1979). We therefore designed a procedure in which the spectral analysis precedes the vowel segmentation. The output records are selected as vowel on the basis of three criteria:

- 0.4-criterium: each data record with F 0 ≤ 200 was rejected (considered)
- high/low ratio (R/L): the definition of low high frequencies records in literature is of uniform (Weinstein et al., 1975) with Le=600 Hz and He=3500-4000 Hz; Kanya & Nakita (1979) with Le=700 Hz and He=8000-30000 Hz, whereas for Dutch speech material Rieveld (1983) define Le=500-622 Hz and He=537-1523 Hz. In the present study we used the filters 1-4 for the low frequency range (250-615 Hz) and the filter 13, 14, and 15 for the high frequency range (2500-4378 Hz), since filter 16 turned out not to be reliable. If the ratio R/L then the date record is rejected as vowel record.

- vocal tract length VT based on the analysis results. Voiceless sounds can also be detected on the basis of VTL per record (Kanka, 1979). VTL can be considered as an approximation to the meaning of the Dutch term "veloformant". VTL criteria needs more work refinement, we obtained average distances for male speaker using the criterion that each vowel record must be at least 3.4 ms longer than that of the following vowel record. In which VTL = 0.4 * D. in which D: is the distance between two records.

In which the automatic vowel segmentation is used on the basis of the following selection criteria is used: 1) the first and last 15% of the recording; 2) a second selection if both previous criteria is used against the remaining records.

The automatic procedure of data processing (manual and automatic) end up in a set of data processing program called Q (Street, 1965) where after recording the analysis, recording, voice and vowel formant values of the files for each speaker have been stored in an output file, together with the ASC value, each time when a record is compared to the previous record on the basis of ASC value, each time when a record is compared to the previous record, the output consists of the final mean values with variance of the parameters mentioned above, and the total number of processed records (for 10 ms voice frames).

The program provides the possibility of manually processing the acoustic system contrast ASC, and of defining the moment when the ASC value stabilizes.

**RESULTS AND CONCLUSIONS**

The results of the manual and the automatic procedure of data processing, one of the main problems was the automatic segmentation from the speech fragment (cf. Kanka, and Nakita, 1979). We therefore designed a procedure in which the spectral analysis precedes the vowel segmentation. The output records are selected as 'vowel' on the basis of three criteria:

- 0.4-criterium: each data record with F0 ≤ 200 was rejected (considered)
- high/low ratio (R/L): the definition of low high frequencies records in literature is of uniform (Weinstein et al., 1975) with Le = 600 Hz and He = 3500-4000 Hz; Kanya & Nakita (1979) with Le = 700 Hz and He = 8000-30000 Hz, whereas for Dutch speech material Rieveld (1983) define Le = 500-622 Hz and He = 537-1523 Hz. In the present study we used the filters 1-4 for the low frequency range (250-615 Hz) and the filter 13, 14, and 15 for the high frequency range (2500-4378 Hz), since filter 16 turned out not to be reliable. If the ratio R/L then the date record is rejected as vowel record.

- vocal tract length VTL based on the analysis results. Voiceless sounds can also be detected on the basis of VTL per record (Kanka, 1979). VTL can be considered as an approximation to the meaning of the Dutch term "veloformant". VTL criteria needs more work refinement, we obtained average distances for male speaker using the criterion that each vowel record must be at least 3.4 ms longer than that of the following vowel record. In which VTL = 0.4 * D. in which D: is the distance between two records.

In which the automatic vowel segmentation is used on the basis of the following selection criteria is used: 1) the first and last 15% of the recording; 2) a second selection if both previous criteria is used against the remaining records.

The automatic procedure of data processing (manual and automatic) end up in a set of data processing program called Q (Street, 1965) where after recording the analysis, recording, voice and vowel formant values of the files for each speaker have been stored in an output file, together with the ASC value, each time when a record is compared to the previous record, the output consists of the final mean values with variance of the parameters mentioned above, and the total number of processed records (for 10 ms voice frames).

The program provides the possibility of manually processing the acoustic system contrast ASC, and of defining the moment when the ASC value stabilizes.