DYNAMIC DETERMINATION OF ACOUSTIC VOWEL CONTRAST

FLORIEN J. KOOPMANS - VAN BEINUM ROB P. DE SAINT AULAIRE

Institute of Phonetic Sciences, University of Amsterdam.

ABSTRACT

A problem in automatic speech recognition as well as in speech synthesis-by-rule is how to cope with the phenomenon of vowel reduction: vowels in connected speech rarely reach their target position (the intended vowel) as defined in isolated-word and isolated-vowel production. This paper describes a semi-automatic dynamic procedure for ongoing vowel analysis, thus providing 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 texts and free conversation) of one 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 but in different speech situations, plays an embarrassing role in speech technology. Vowels in connected speech rarely reach their target position (the intended phoneme) as defined in isolated-word and isolated-vowel production. In speech synthesis we badly need a model to describe this variability in order to increase intelligibility as well as naturalness, whereas in automatic speech recognition vowel reduction is an annoying phenomenon that we do not know how to cope with.
It is known that the degree of acoustic contrast between the vowels in a speaker's vowel system is dependent on various factors, partly global and partly local, but it is not clear as to how far all these factors are mutually dependent or independent.
In literature we can find, apart from socio-phonetic and linguistic factors, a large number of acoustic-phonetic factors that are believed to be responsible for the variability and the reduction of acoustic vowel contrast (for a detailed overview see Koopmans-van Beinum, 1980).
As far as the acoustic-phonetic factors are concerned (like speech rate, stress, intonation, 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 these factors and more specifically their hierarchical structure have been studied only fragmentarily yet. Lindblom (1963) for instance postulates that duration is the main determinant of vowel reduction, whereas Delattre (1969) claims stress and speech rate to be primary determinants with duration as a product of stress and tempo and therefore a secondary determinant. Gay (1977) and Den Os (1985) both show that an increase of speech rate not necessarily affects the formant frequencies of the vowels. Furthermore Koopmans-van Beinum (1980) indicates a different relation between stress and vowel duration for read texts as compared to texts with a free choice of words (retold story or free conversation). Also from perceptual studies on stress (e.g. Van Katwijk, 1974; Rietveld, 1983; Rietveld and Koopmans-van Beinum, to appear) the relation between loudness, intonation, speech rate, and vowel contrast reduction turns out to be a very complicated one.
In order to reach a better understanding of the relations and the hierarchical structure of the great variability of vowels, it is deemed necessary in our approach of the speech signal to make a distinction between 'global' factors (socio-phonetic aspects such as speaker, speech situation, and sex) influencing this variability, and 'local' factors (acoustic-phonetic and linguistic aspects within the neighbouring context).
We therefore started a project in order to develop and apply strategies to make optimal use of acoustic, socio-phonetic, and if possible also of linguistic information with respect to the variability in the realization of vowel phonemes. This will be done by means of a semi-automatic method for dynamic vowel analysis and cumulative data processing in three phases:
a) Any speech fragment of any speaker may be subjected to a dynamic acoustic-phonetic analysis to provide information on global aspects as mentioned above about the present vowel system (sex of the speaker, overall speech rate, degree of vowel contrast, etc.). Moreover the acoustic parameter values in the dynamic vowel analysis provide the possibility to define the moment when the global measure for acoustic system contrast (ASC) stabilizes. This indicates the duration of the speech sample needed for defining this value (and other global measures), and for dynamically adjusting it, if use is made of a moving window.
b) Subsequently local measures of acoustic vowel contrast or degree of reduction and variability will be developed based on acoustic-phonetic parameters as fundamental frequency, formant frequencies, bandfilter values, vowel duration, amplitude.
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 acoustic parameter values combined with global contrast measures and general information on the present vowel system, and defining the hierarchical structure of factors influencing the variability in vowel phonemes.
This paper reports on our first steps within this project, viz. the development of a method for the dynamic determination of the global measure for acoustic system contrast and its application to two quite distinct languages, Dutch and Japanese.
So three main questions have to be answered: 1) what differences yields the dynamic cumulative analysis method compared to the traditional static one; 2) what differences yields the (semi-)automatic procedure compared to the traditional manual one; 3) is the dynamic (semi-)automatic procedure applicable to two phonetically quite distinct languages.

DESIGN OF A DYNAMIC ANALYSIS AND DATA PROCESSING METHOD

As the aim of the present subproject was to develop a (semi-)automatic dynamic procedure of data processing, two parallel methods had to be compared: a) the traditional method making use of manual segmentation of vowels in the digitized speech fragment by means of a speech editor, followed by a dynamic acoustic-phonetic vowel analysis, and b) a (semi-)automatic method by carrying out a dynamic acoustic-phonetic analysis firstly on all speech frames, followed by an automatic vowel segmentation. Subsequently both methods are followed by a data processing program (based on formant frequencies or based on bandfilter values) which calculates in a cumulative way the acoustic system contrast measure ASC (Koopmans-van Beinum, 1980; De Saint Aulaire, 1986). This ASC measure is defined by the total variance of all vowels in the present vowel system, based on frequencies of the first (F1) and second formant (F2), (transformed in 100 * 10log Hz), using the formula:

$$ASC = \frac{1}{N}\sum_{j=1}^{N}(\vec{V}_j - \vec{C})^2 \quad \text{in which}$$

$\vec{V}_j$ = the 2-dimensional vector of vowel j in the F1/F2-plane,
$\vec{C}$ = the 2-dimensional vector of the centroid C,
N = the number of vowels in the vowel system.
Apart from our formant-based ASC-measure we also developed a similar ASC-measure based on bandfilter variance, which turned out to be a good alternative for the formant-based one, but in this paper we left it out of consideration (Koopmans-van Beinum and De Saint Aulaire, 1986).

SPEECH MATERIAL

It is claimed that in so-called syllable-timed 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-timed languages like e.g. English, Russian, and Dutch. In previous work, however, we met for Dutch and for Japanese a similar degree of vowel reduction expressed in comparable values of acoustic system contrast (De Graaf & Koopmans-van Beinum, 1982/83). Therefore we decided to test our dynamic analysis and data processing method on Dutch as well as on Japanese speech material.
The Dutch vowel system consists of twelve more or less monophthongal vowels and three diphthongs. All vowels and diphthongs may occur in stressed as well as in unstressed position. Furthermore about 30% of all vowel phonemes in Dutch consists of schwa 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 schwa are short, at least in connected speech. Spectrally the Dutch short vowels are not more centralized than the long vowels (for more details see Koopmans-van Beinum, 1980).
The Japanese vowel system is a 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 are pronounced without serious qualitative distortion. However, the incorrectness of the latter claim is proved by De Graaf & Koopmans-van Beinum (1982/83; 1984) who demonstrated a similar degree of reduction in connected speech for a number of languages including Japanese.
As for Dutch we used recorded speech material of the same trained male speaker as in Koopmans-van Beinum (1980). This provided us with the possibility to compare the results of the present procedures with previous results. Nevertheless an important difference remained: in the present speech material we used all segments automatically labelled as being vowel-like in the chosen speech fragment, and also in the order in which they occurred. Moreover measurements were carried out dynamically with ten millisecond steps. This means that frequency of occurrence of all vowels in normal running speech now got the attention it deserves, and that the duration of each occurring vowel weighs proportionally in the calculation of the acoustic system contrast. In the former study ten items of each vowel were used and were measured only at one point more or less in the middle of the vowel. An accidental advantage of the present 'weighing' procedure is the fact that it is no longer necessary to 'label' the vowel segments, i.e. we no longer need to know which vowels the speaker intended to say. The acoustic system contrast ASC of a speech fragment of a specific speaker is defined now by the total variance of all vowels, i.e. of all analysed 10 ms vowel frames, just as they occur in the speech fragment. The moment at which this ASC stabilizes actually defines the length of the speech fragment

needed for the determination of the ASC for that specific speaker in that specific speech situation. As to how far length of fragment depends on speaker, on speach situation, and on language is one of the research questions of the project as a whole.
As for the Dutch speaker we made use of two speech situations: free conversation (a 30 sec fragment) and read text (a 10 sec fragment). The speech fragments were selected from existing recordings, which provides us with the possibility to compare the static and dynamic analysis method using fragments of 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 spectral energy stabilize within that period of time (Li, Hughes, and House, 1969; Zahorian and Rothenberg, 1981). Our choice of only a 10 sec fragment of read text is defined by the results obtained from the free conversation fragment and the need of confining the material.
As for Japanese the speech material of one male speaker (free conversation 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 did 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, only manual analysis), Dutch read text (10 sec, manual and automatic), Japanese conversation (10 sec, only automatic) and Japanese read text (10 sec, manual and automatic).

MEASUREMENTS

By means of a speech editing program (Buiting, 1981) all vowel items in the digitized speech fragments were isolated in such a way that the starting-point of the vowel was considered to be the place were the formant pattern of the vowel was clearly visible in the oscillogram for the first time, and the end was taken to be the point were the specific formant pattern disappeared. In case of adjacent voiced consonants only those successive samples were segmented that did not display any auditorily nor visually observable consonant information. Once the vowel segments were isolated, their durations were of course known as well. From the 30 sec fragment of free conversation 121 vowels could be selected with an average duration of 68.66 ms. From the 10 sec fragment of read text 57 vowels were segmented with an average duration of 71.12 ms. Each vowel segment has been analysed dynamically in 10 ms steps (window size 25.6 ms) by means of a spectral analysis program called QQ (Weenink, 1986) using an LPC order of 12 as a standard.
Apart from a number of other data, not relevant for this study, the program QQ provides us with:

- fundamental frequency (F0) using the sieve algorithm (Duifhuis et al., 1982);
- formant frequencies determined by some optional methods; in our case we used Prony's method for LPC-analysis;
- bandfilter values: a bandfilter analysis based on the FFT amplitude spectrum is carried out with filter specifications given by Sekey and Hanson (1984).

The resulting data are stored in analysis files consisting of successive records, each of them containing the analysis results of one 10-ms vowel frame. In this way all kind of selections and calculations can be carried out in subsequent data processing programs.
With respect to the development of an automatic procedure of data processing, one of the main problems to overcome is the segmentation of vowels from the speech fragment (cf. Kasuya and Wakita, 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:

- F0-criterium: each data record with F0=0 was rejected (unvoiced);
- high/low ratio (H/L): the definition of low and high frequency areas in literature is not uniform: Weinstein et al. (1975) use L=0-900 Hz and H=3700-5000 Hz; Kasuya & Wakita (1979) use L=0-500 Hz and H=3800-5000 Hz, whereas for Dutch speech material Rietveld (1983) defines L=262-2230 Hz and H=5575-11150 Hz. In the present study we used the filters 1-6 for the low frequency area (92-856 Hz) and the filters 13, 14, and 15 for the high frequency area (2549-4239 Hz), since filter 16 turned out not to be reliable in all cases. So if the ratio H/L>1 then the data record is rejected as a vowel record.
- vocal tract length VTL: based on the analysis results of QQ this program calculates also the VTL per record (Wakita, 1977). Considering the formant frequencies and VTL together revealed that in case of low (nasal) F1 the VTL showed very unreal values (0.0 or -1.0 cm), whereas for records with an extreme high F1 value (e.g. F1>1500 Hz) the calculated VLT attained to about 10 cm. All other records display a more or less normal distribution of VTL values. Although this VTL criterium needs some more refinement, we obtained satisfactory results in this study by using the criterium that each vowel record had to attain a VTL value of:

$$\overline{VTL} - 0.5*s.d. \leq VTLx \leq \overline{VTL} + 1.0*s.d.$$

in which VTLx = the VTL of data record x.
Within the automatic vowel segmentation program the following hierarchy of criteria is used: 1) a first selection is done based on the F0- and the high/low ratio criterium; 2) a second selection is done based on the VTL-criterium, applied to the remaining records.
Both procedures (manual and automatic) end up in a set of data processing programs calculating cumulatively (i.e. record after record) the mean values and the variance of the fundamental frequency, of the first four LP-formants, and of the 16 bandfilter values. During processing the mean F1, F2, mean bandwidths of each formant, mean level of each bandfilter, F0, and the ASC are stored in an output file, together with the deviation of the new ASC compared to the preceding ASC value, each time when a record is closed. At the end of the processing the output consists of the final mean values with variance of the parameters mentioned above, and the total number of processed records (= the number of 10 ms vowel frames).
The program provides the possibility of cumulatively 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, compared with the results from our preceding studies, are described in detail in Koopmans-van Beinum and De Saint Aulaire, 1986 and will be presented at the Congress. Summarizing the results we can state that
1) the measure for acoustic system contrast ASC, cumulatively processed on the output data of a dynamic acoustic-phonetic vowel analysis, compares favourably with the ASC-values as processed on output data from static vowel analysis; it should be kept in mind, however, that in the dynamic method all vowel segments (including diphthongs and schwa) are processed in their total duration;
2) the here presented automatic procedure for processing running speech provides us for the time being with a satisfying possibility to define quickly and for extended speech material, global reduction data in terms of acoustic system contrast. Moreover our methods provide tools to recognize where discontinuities in the global measures occur in the processed speech material, possibly caused by accentuation and indicating important events in running speech.
For the analysed Japanese speech material the resulting ASC-values fit well with the previous results of the static analysis method on speech material of three other speakers. Nevertheless the manual and automatic procedure, applied on the Japanese read text, display slight differences in the results, possibly caused by our poor knowledge of the Japanese language necessary for proper manual segmentation.
For Dutch as well as for Japanese free conversation the measure for acoustic system contrast stabilizes within 2 sec. of vowel material, which means for both languages that about 6 sec. of free conversation should be enough. The high number of schwa sounds in Dutch will cause a lower ASC-value than in languages without schwa phonemes. In free conversation this is confirmed in our data, but more speakers are needed to prove the reliability, since vowel reduction turns out to be greatly speaker dependent. In the read texts of both languages, the fluctuations in ASC-value are much more persistent, mainly caused by F2 fluctuations which are more violent for Japanese than for Dutch. Here again we can possibly trace the influence of the frequently occurring schwa sounds in Dutch.
In the near future our research will concentrate on combining the dynamically processed global measure for acoustic system contrast with local parameter values in order to be able to control the influence of vowel reduction aspects on speech recognition and speech synthesis.

REFERENCES

Buiting, H.J.A.G. (1981). SESAM, Speech Editing System Amsterdam, IFA-report 70, Amsterdam.
Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. IRAL 7, 295-325.
Duifhuis, H., Willems, L.F. & Sluyter, R.J. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. J. Acoust. Soc. Am. 71, 1568-1580.
Gay, T. (1977). Effect of speaking rate on vowel formant movements. Haskins Lab. Status Report on Speech Research SR-51/52, 101-117.
Graaf, T. de & Koopmans-van Beinum, F.J. (1982/83). Vowel contrast reduction in Japanese compared to Dutch. IFA-Proceedings 7, 27-38.
Graaf, T. de & Koopmans-van Beinum, F.J. (1984). Vowel contrast reduction in terms of acoustic system contrast in various languages. IFA-Proceedings 8, 41-53.
Kasuya, H. & Wakita, H. (1979). An approach to segmenting speech into vowel-like and nonvowel-like intervals. IEEE Trans. ASSP-27, 319-327.
Katwijk, A.F.V. van (1974). Accentuation in Dutch. Diss. RU Utrecht.
Koopmans-van Beinum, F.J. (1980). Vowel Contrast Reduction. Diss. Univ. of Amsterdam.
Koopmans-van Beinum, F.J. & Saint Aulaire, R.P. de (1986). A method for the dynamic determination of acoustic vowel contrast. IFA-Proceedings 10, 1-17.
Li, K.P., Hughes, G.W. & House, A.S. (1969). Correlation characteristics and dimensionality of speech spectra. J. Acoust. Soc. Am. 46, 1019-1025.
Lindblom, B.E.F. (1963). Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.
Os, E.A. den (1985). Vowel reduction in Italian and Dutch. PRIPU 10-2, 3-12.
Rietveld, A.C.M. (1983). Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands. Diss. KU Nijmegen.
Rietveld, A.C.M. en Koopmans-van Beinum, F.J. (to appear). Vowel reduction and stress. Speech Communication.
Saint Aulaire, R.P. de (1986). Klinkerreductie en taalstructuur: een verwerkingsmethode. IFA-report 85, Amsterdam.
Sekey, A. & Hanson, B.A. (1984). Improved 1-Bark bandwith auditory filter. J. Acoust. Soc. Am. 75, 1902-1904.
Takebayashi, S. (1975). The vowels of Japanese and English. Lexicon 4, 49-67.
Wakita, H. (1977). Normalisation of vowels by vocal tract length and its application to vowel identification. IEEE Trans. ASSP-25, 183-192.
Weenink, D.J.M. (1986). QQ: een programma voor analyse, resynthese en herkenning van klinker-segmenten. IFA-report 82, Amsterdam.
Weinstein, C.J., Mc Candless, S.S., Mondshein, L.F. & Zue, V.W. (1975). A system for acoustic-phonetic analysis of continuous speech. IEEE Trans. ASSP-23, 54-67.
Zahorian, S.A. & Rothenberg, W. (1981). Principal-component analysis for low-redundancy encoding of speech spectra. J. Acoust. Soc. Am. 69, 832-845.