Within-vowels correlation in speech and speaker recognition

Stefan Grocholewski
Institute of Computing Science
Poznan University of Technology, Poland

Introduction

The idea presented in this paper consists in considering, instead of (or in addition to) direct measurements, the correlation of the measurements between pairs of vowels for speech and speaker recognition. The correlation is defined as the F-ratios, i.e. ratios of between – to within – speech vector parameters variation. In automatic speech recognition, analysis of correlation between words in an utterance was studied, e.g., in [1]. In the appropriate experiments, the scores were improved by several percent. In the domain of automatic speaker recognition, the values of F in ANOVA were used in estimating selected parameters in the word OKAY for verifying talkers [2].
In speech recognition, the philosophy of determining significant features that are independent of the speaker centers on the search for large values of F.
For speaker recognition purposes we intend to find the speaker dependent features coefficients for which the values of F have typical large dispersion
In the present investigation, cepstral coefficients are used to find invariant features for the specific purpose of dichotomous classification. This being just a pilot study, we have limited ourselves to features that distinguish pairs of vowels. It should be noted that a similar task could be defined in the formant domain. The essential difference is, however, that while an unambiguous determination of formant frequencies does not exist, there is no problem of ambiguity in determining the cepstral coefficients. We abstract here, for the present, from problems of noise, channel characteristics, etc.

Correlation analysis for speaker recognition

The purpose of the work is to study if the use of within-vowels correlation can improve the performance of speaker verification systems. An understanding of the acoustic properties, as well as the nature of within- and between speaker variation is of great importance in speaker recognition. For example, formant frequencies are closely dependent on the vocal tract length. This results in a strong relation not only between the formants in a certain vowel but also between different vowels spoken by the same speaker. The first finding, the relation between the formants in the same vowel, is very often used in speaker normalization procedures in speech recognition systems. The second one, the relation between different vowels has been rather used for speech recognition, see Blomberg [1] or Niyogi [3]. In the pilot study we have observed the different correlation between the same pairs of phonemes among different speakers. Unlike in other studies, instead of using the phonetic approach consisting in considering the formant analysis, we have used the cepstral coefficients analysis. The reason was the problem with the exact determination of formants in real applications.
The proposed idea is illustrated in the Fig 1.

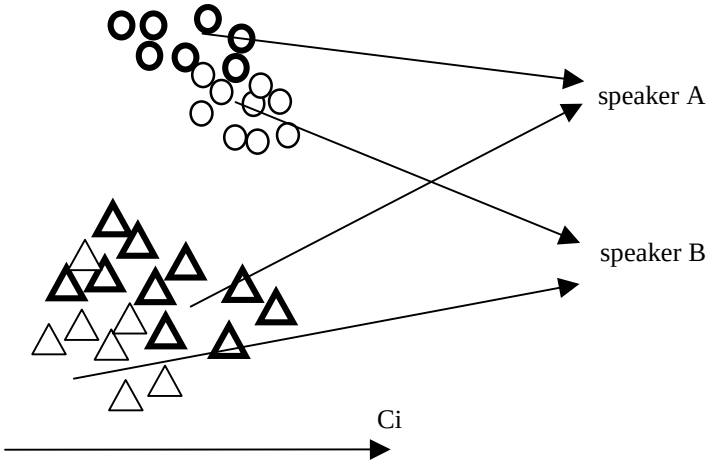

Fig. 1. Illustration of the idea of the proposed method

Fig. 1 presents the axis of the parameter Ci and observations which belong to two classes: circles and triangles. The observations are from two speakers represented by a thick line (speaker A) and a thin line (speaker B) .
It is impossible to identify the speakers reliably using the parameter Ci only from one class of observations. We can find the same values for both speakers. Our approach consists in using the value of correlation between observations

from two classes. We can conclude that correlation between observations represented by thick lines (speaker A) is much higher than between observations represented by thin lines (speaker B).
We assume that different dimensions possess different abilities to distinguish speakers or to distinguish observations.
We intend to investigate the use of the values of F-ratios for speaker verification.

Experiment design

We have used the data from our speech database CORPORA [4]. Ten male speakers have been selected. For each speaker we have chosen the middle parts of 600 vowels (6 Polish vowels x 100 items in different contexts) from two separated parts of CORPORA.
The data (6000 vowels) have been divided into two subsets to study the dependence on the learning sets. A series of univariate ANOVAs was performed to calculate the F-ratio for each cepstral coefficient [5].

Acoustic analysis

Let us consider the F-ratios for 12 cepstral coefficients in the case of the pair {*a*, *i*} for 10 speakers.

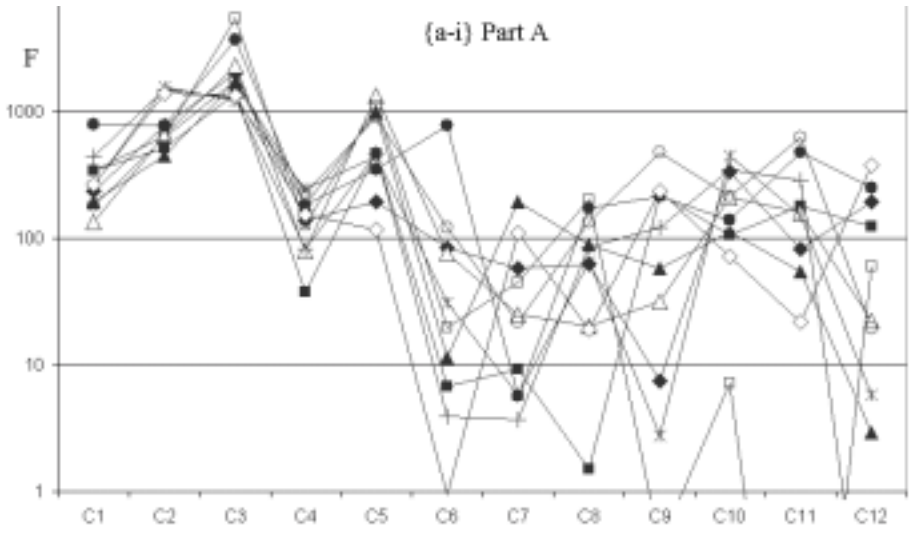

Fig. 2. F-ratios for 12 cepstral coefficients
of the pair {*a*, *i*} for ten speakers.

Analyzing Fig. 2 we can conclude that for all speakers the F-ratios for C2 and C3 are very high (the coefficients are not correlated and are not suitable for speaker recognition). From the speaker recognition point of view we should consider C6, C8, C9 or C12 because they depend on the speaker.
To see whether the F-ratios depend on the learning set or not, Figures 3 and 4 present the detailed comparison of F-ratios for two speakers: *ao* (Fig. 3) and *dg* (Fig. 4), for the same pair of vowels {*a*, *i*} but for different sets.

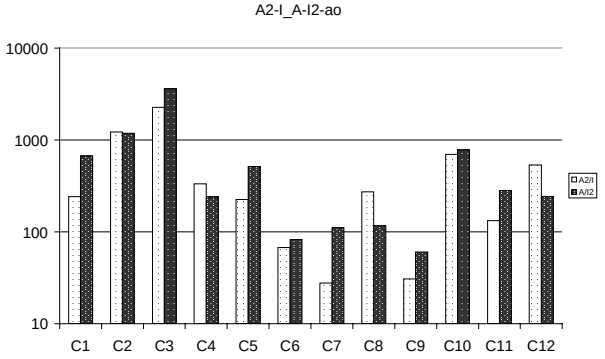

Fig. 3. Comparison of F-ratios for two learning data –speaker *ao*

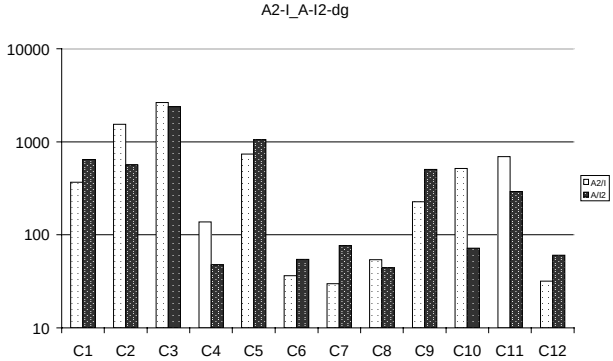

Fig. 4. Comparison of F-ratios for two learning data –speaker *dg*

Comparing Fig. 3 and 4 we can see the dependence of the F-ratios on the learning sets as well as the differences for both speakers (C8, C9, C12). The conclusion is that the presented idea can be suitable as the supplementary method for speaker recognition

An analysis of variability in the cepstral domain

The problem of finding invariant features in the speech signal is known ever since the emergence of acoustic phonetics. In the beginning, it referred to finding features efficient in an unambiguous identification of certain linguistic units, such as vowels. It was found quite soon that a fairly good classification of vowels could be based in the values on the two lowest formants [6]. It transpired, however, that that formant spaces of different vowels may overlap partly if spoken by different speakers, especially speakers of different genders. It follows that the values of the formants themselves cannot, after all, be considered the invariant features of vowel phonemes.

We intend to use the methodology based on correlation analysis for speech recognition. In the present study, the variability of cepstral coefficients was investigated using an analysis of the correlation between the appropriate cepstral coefficients in pairs of vowels. In Fig. 2, a correlation is described in terms of the coefficient F for 12 cepstral coefficients in every pair of vowels {a-i}. In the Figure, those coefficients for which the values of F have typical small dispersion (C_1...C_5) as against those where the dispersion is large (C_6...C_{12}). The former are independent of the speaker (within the sample under investigation), and of the context. The latter are dependent on the context and/or the speaker.

As can be seen in Fig. 2, in the case of the pair {a-i}, the least correlated coefficients are C_1...C_5, with the lowest correlation for C_3. We should remark that the figure refers to a pair which are very different acoustically, and are easily distinguishable.

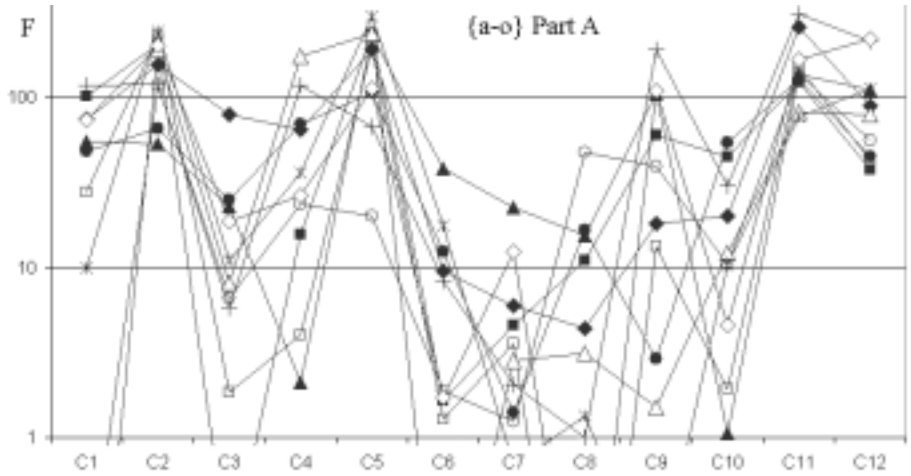

Fig. 5. The F values for the pair {*a-o*} - part A of the database

Fig. 5 shows the pair {o-a} which are contingent in the F1 x F2 plane. As in the foregoing pair, here, too, cepstral coefficients can be found that are essential, this time for the pair {o-a}. It is interesting to note that for all speakers, the smallest correlation is that of C_{11}.

In order to test the predictability of the results, the B corpus was used, and the results are given in Fig. 6, analogous to Fig. 5. The differences are assumed to be due to differences in phonetic environment, but there is an essential replicability of the cepstral coefficients. Juang and Rabiner [7] showed that the variation of higher quefrency terms in cepstral domain is due to inherent artifacts of the analysis procedure. In contrast the variation of lower quefrency terms is primarily due to variations in transmission channel, vocal tract and speaker's characteristics.

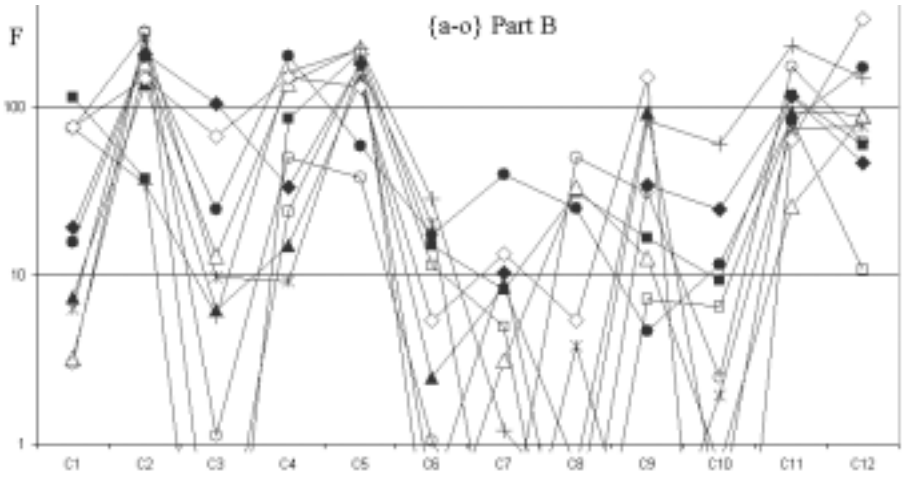

Fig. 6. The F values for the pair {*a-o*} - part B of the database

The experiments were carried out for 10 male speakers in the database CORPORA. Using a search function, 100 realizations of each vowel as pronounced by the 10 male speakers for a total of 6000 vowel tokens. This total was divided into two parts, A and B. The different vocabulary of the two parts was expected to ensure equal effect of context variability in the two corpora. Table 1 shows the results of an analysis of correlation for all 15 pairs of phonemes in part A. For each pair the cepstral coefficients are given that are singled out by the lowest correlation

Pair	Cepstral coefficients
i-y	C_1, C_2, C_6
i-e	C_1, C_2, C_3, C_8, C_9
i-a	C_1, C_2, C_3, C_4, C_5
i-o	C_1, C_2, C_3, C_4
i-u	C_1, C_3, C_4, C_5
y-e	C_2, C_8
y-a	C_2, C_3, C_4, C_5, C_6
y-o	C_1, C_3, C_4, C_6

Pair	Cepstral coefficients
y-u	C_1, C_2, C_3, C_6
e-a	C_3, C_5, C_6
e-o	C_4, C_6, C_8
e-u	C_1, C_2, C_6, C_8
a-o	C_2, C_5, C_{11}, C_{12}
a-u	C_2, C_3
o-u	C_3, C_8

Table 1. The cepstral coefficients that are significant for the recognition of vowels within pairs.

Conclusions

In the paper the idea of using the within-vowels correlation has been described. The pilot experiments showed the this idea is suitable as the supplementary method for speaker recognition. In the speech recognition domain rather than looking for universal features in all the vowels, pair of vowel classes were subjected to analysis. Produced by 10 different speakers, the 15 pairs of the Polish vowel classes were defined in terms of cepstral coefficients, assumed to be the most useful in a dichotomous classification. The selected criterion was the value of the correlation of the respective cepstral coefficients for each pair of vowel types.

References

[1] Blomberg M., *Within-utterance correlation for speech recognition*, European Conference on Speech Communication and Technology, EUROSPEECH'97, Rhodes, Greece, 1997, pp.2479-2482.

[2] Eliot J., *Auditory and F-pattern variations in a australian OKAY*, Eight Australian Int.Conf. on Speech Science and Technology, Canberra, 2000, pp.148-153.

[3] Niyogi P., Zue V., *Correlation Analysis of Vowels and their Application to Speech Recognition*, Proceedings of Eurospeech'91, p. 1253 –1256.

[4] Grocholewski S., *CORPORA –Speech Database for Polish Diphones*, Proceedings of Eurospeech'97, p.1735-1738

[5] Grocholewski S., *Within vowels correlation for speaker recognition*, Int. Conf. on Systemics, Cybernetics and Informatics,. SCI'2001, Orlando, USA, 2001, pp. 384-387.

[6] Fant G., *Acoustic theory of speech production*, Mouton & Co., S-Gravenhage, 1960.

[7] Juang B.H., Rabiner L.R., *On the use of bandpass liftering in speech recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-35, no. 7, July 1987, pp.947-954.