Elements of speaker variability in some voiceless phonemes

Ewa Łukasik

Institute of Computing Science, Poznań University of Technology
ul. Piotrowo 3a, 60-965 Poznań, Poland
e-mail: lukasik@put.poznan.pl

Abstract. The paper presents results of the analysis of speaker variability elements in some voiceless phonemes, namely voiceless stop consonants. Gender and speaker recognition has been performed using the ANN with context independent cepstral and wavelet packet features. The gender recognition accuracy reaches 80% and is rather high for non-pitched phonemes. Speaker identification accuracy was significant only in training phase.

1. Introduction

Speaker variability, such as gender, accent, age, speech rate and phonemes realizations, is one of the main difficulties in speech signals recognition. Some speaker adaptation methods are used in speaker independent recognition systems, e.g. reflecting intrinsic characteristics about specific speakers by re-training the system using appropriate corpora or building multiple models of smaller variances, such as gender dependent model [1].

This paper presents results of analysis of speaker variability elements in voiceless phonemes. The motivation for this work arose during author's research on context and speaker independent recognition of voiceless stop consonants using new methods of time frequency analysis. ([2][3]). Strong gender dependence as well as some speaker impact on recognition results has been observed confirming conclusions reported elsewhere, e.g. in [4]. How big this influence is in terms of classification rate seemed to be interesting to explore in contrast to common conviction that the most significant feature of gender is pitch [1].

Voiceless stop consonants (/p/, /t/, /k/) are unique elementary speech signal categories (phonemes) of non-stationary character. Their characteristics are formed by the place of articulation. Speaker dependent information comes mainly from the vocal tract. The excitation generated by the airflow from lungs, in case of voiceless stops has the form of compression. It results from the release of a completely closed and pressurized vocal tract, therefore the excitation place is inside the vocal tract itself [5] and its role in determining speaker dependence is enhanced.

The paper focuses on the investigation of gender information and speaker identification in voiceless stops. Phonemes have been extracted from speech database CORPORA [6] for several speakers of both genders.

2. Speech data and features

Voiceless stops used for experiments have been extracted from speech database CORPORA [6] - each phoneme from different, unrepeatable utterance. The analyzed segments are as long as extracted phoneme from 80 to over 1200 samples (5 ms to almost 100 ms at 16kHz sampling rate). Two subsets of speech data have been created: STOP 1 – with utterances of 2 men (95/p/, 98/t/, 123/k/), and 2 women (129 /p/, 248 /t/, 161 /k/) where only carefully pronounced phonemes have been kept and STOP 2 – with no pre-selection - of 5 men (354 /p/, 644 /t/, 538 /k/) and 5 women (425 /p/, 774 /t/ 642 /k/).

Two independent sets of features have been used: a set of 20 conventional cepstral coefficients and parameters based on Shannon entropy matrices calculated from wavelet packet coefficients, selected by means of Singular Value Decomposition (a modified method used in [2]). Classification has been carried out using ANN with one hidden layer and backpropagation learning algorithm. 70%.

3. Classification results and discussion

Two experiments have been carried out: for gender and speaker identification. In Table 1 exemplary classification results are presented for two feature sets: one based on wavelet packets coefficients and the other on standard cepstral coefficients. The results for data in STOP 1 and STOP 2 are presented separately for individual phonemes. The classifier was the multilayer perceptron neural network with backpropagation training algorithm (one hidden layer). In training phase 70% of input vectors have been used, the rest being used for testing. Division into training and testing sets has been performed randomly.

Table 1. Gender classification rate (in %) for two sets of stop consonants /p/, /t/, /k/ with cepstral and wavelet packets entropy as attributes

Phoneme		Cepstral coeff.		Wavelet packets	
		Male	Female	Male	Female
/p/	STOP 1	79,31	82,93	77,42	92,31
	STOP 2	64,42	63,49	65,81	67,10
/t/	STOP 1	96,00	97,65	85,71	97,62
	STOP 2	73,10	79,33	73,34	71,64
/k/	STOP 1	74,36	80,39	66,67	73,89
	STOP 2	79,17	78,02	73,95	74,55

The results presented in Table 1 show that there is substantial information concerning gender in voiceless plosives. The average classification rate is about 80%. The most gender distinctive is plosive /t/. The classification rate for STOP 1 containing well selected utterances is higher than for STOP 2 where no attention has been paid for the quality of speech material. Number of wavelet packet coefficients indicated by SVD method was much smaller than that of cepstral coefficients and it might be the reason for a difference in classification rate for both methods.

Second experiment concerned speaker identification. 20 cepstral coefficients have been used as input features for ANN classifier. There were a substantial diiference in results of training and testing phases. Neural network was able to learn the characteristics of input data giving up to 100% of correctness (for /t/) for well extracted phonemes in data set STOP1 (4 speakers). New data provided in the testing phase appeared so different in comparison with the material with which the ANN was trained that the averaged classification rate dropped down to about 65% in average. It is interesting to note that big differences have been observed for individual speakers (from 40% to 80 % of accurate identification). As might be expected, the results for the second data set STOP2 have been similar with the overall smaller classification accuracy. In both training and testing phases results for individual speakers and phonemes were very different, reaching 65% and 45% respectively on the average. In Fig.1 we present the averaged results for both aspects of speaker variability analysis, namely gender and speaker identification accuracy in training and testing phases for data sets STOP1 and STOP2.

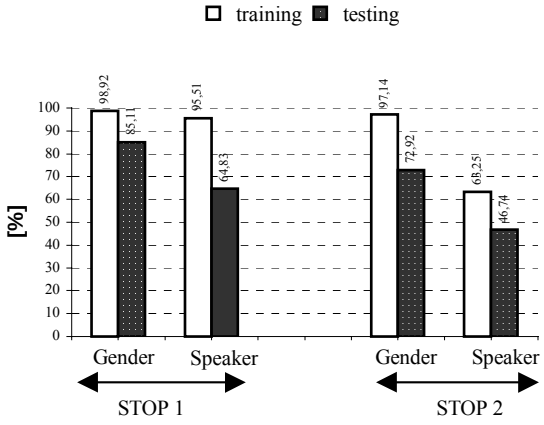

Fig. 1. Comparison of gender and speaker identification rate it training and testing phases for two data sets of stop consonants: STOP1 and STOP2 with cepstral coefficients as attributes

The results obtained in the experiments are to certain extent surprising. The low energy non pitched phonemes give substantial information concerning gender. It means that they really reflect the characteristics of the vocal tract of male and female speakers. Since quite big differences have been observed in the first cepstral coefficient, the distnction may come from differences of phoneme energy while uttered by men and women. Some vowel context may have also impact on the recognition [7]. However the reflection of vocal tract in voiceless phonemes is not big enough to provide reliable cues for speaker identification.

4. Conclusions

In the paper some investigations concerning speaker variability in voiceless plosive consonants (/p/, /t/, /k/) have been presented. We concentrated on gender and speaker identification, using speech material from CORPORA - the speech database of Polish. We tried to answer the question how big is this influence is in the case of low energy non pitched phonemes, especially because commonly pitch related features are used for speaker identification.

Cepstral coefficients constituted main set of attributes, wavelet packets being used principally for confirmation of the results. The gender identification accuracy using ANN classifier were astonishingly high and exceeded 75% on average. The main cue may be in the difference of energy in male and female utterances. Some other elements of vocal tract characteristics may also be significant. However the attributes are not distinctive enough to give the reliable tips for speaker identification.

6. References

[1] C. Huang, T. Chen, S. Li, E. Chang, J. Zhou, "Analysis of Speaker Variability", *Proc. Eurospeech 2001*, Aalborg, Denmark.

[2] E. Łukasik, "Wavelet Packets Based Features Selection for Voiceless Plosives Classification", *Proc. ICASSP*, Istambul 2000, pp.689-692.

[3] E. Łukasik, "Classification of Non-Stationary Acoustic Signals Using Wavelet Packet Based Features", *Proc. XXII National Conf. on Circuit Theory and Electronic Networks*, 1999, pp. 491-496.

[4] K.N. Stevens, S.Y. Manuel, M. Matthies, "Revisiting Place of Articulation Measures for Stop Consonants: Implications for Models of Consonant Production", *Proc. of Int. Congress of Phonetic Science*, San Francisco 1999, pp. 1117-1120.

[5] J.B. Campbell, Jr., “Speaker Recognition: a Tutorial”, Proc. of the IEEE, vol. 85, no 9, 1997, pp. 1437-1462.

[6] S. Grocholewski, "CORPORA - Speech Database for Polish Difones", *Proc. Eurospeech'97*, Rhodes, Greece, 1997, pp. 1735-1738.

[7] P.J.B. Jackson, "Acoustic cues of voiced and voiceless plosives for determining place of articulation", *Proc. of CRAC Workshop*, Aalborg, Denmark 2001.