ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2010. No. 9(105)
ELEKTRONIKA IR ELEKTROTECHNIKA

SIGNAL TECHNOLOGY
T 121
SIGNALŲ TECHNOLOGIJA

Analysis of Factors Influencing Accuracy of Speech Recognition

G. Čeidaitė
Department of System Analysis, Vytautas Magnus University,
Vileikos str. 8, LT-44404 Kaunas, Lithuania, e-mail: g.ceidaite@ist.vdu.lt
L. Telksnys
Institute of Mathematics and Informatics,
A. Goštauto str. 12, LT-01108 Vilnius, Lithuania, e-mail: telksnys@ktl.mii.lt

Introduction

The accuracy of the recognition data is important problem of speech recognition system. The main attention is given to algorithms, features and templates. But no less important are paralinguistic factors such us speaker's age and sex, physiological and psychological status, voice timbre, environment (e.g. bus station, metro, cafe, school, etc.), signal channels, background noise, and others [1, 7].

Human voice timbre, physiological status, psychological status and background noise are the main Para-linguistic factors which influence is analyzed in speech recognition.

Various recognition methods are used to analyze influence on speaker's emotional state. One of them is to try to recognize when speaker said word in a joyful, upset, surprised or other tone [8, 12]. When analyzing influence of the environment the method when near recognition system is playing different noises is used [2].

Another very important object for recognition is microphone. Here different methods are also used and one of them is when two different microphones are used, with one microphone for the learning system, and another one for the recognition [3]. Examples of other systems could be changing distance between speakers and recognition systems [4, 9], and using more than two microphones and background noise [5, 6].

The aim of this paper was to find out through the experimental analysis what direct influence do the Para-linguistic factors found in natural environment have on the recognition of speech signals. The main attention was given to the specific influences such as speakers' contingent, environment, training conditions and features.
The results of the research are given at the end of the paper.

Statement of the problem

The accuracy of speech recognition was tested with these factors:

- features of speech commands used in recognition;
- modeling of speech command etalons used in the recognition;
- characteristic of speech signals used to create etalons;
- environments' in which speech commands' recognition are used;
- contingent of speakers used in the speech recognition.

In the experiment used speech commands recognizer was based on the dynamic time warping method [13].

Solution of the problem

A lot of analyses of speech commands recognition problems are done in the laboratories [10,11]. However, people using recognition systems are working in different environments with different characteristics from those the systems have been developed in. Thus, while analyzing the practical use of speech command recognition technologies, we had to explore the factors that influence accuracy of these technologies. Therefore the experiment was concentrated to those problems.

Influence of the factors on the accuracy of speech command recognition was examined in this way. At first the situation was examined, then speech command etalons were created in the four different environments and by one woman-speaker.

Three types of features were used:

- Cepstrum coefficients;
- Central cepstrum coefficients;
- Coefficients of speech signal autoregressive process.

To create speech command etalons were used two etalon types:

The first etalon type – one speech signal command etalon is selected from the set of two speech signal commands.

The second etalon type– to create this speech signal command etalon five speech signal commands are selected from the set of six speech signal commands.

In Fig.1 environments used for the creation of speech signal command etalons and for making experiment of speech signal recognition are presented. More details about environments:

Environment A was a workplace with four computers. Also there were printer, shelves, two windows, settee, coffee-table, hanging and standing shelves.

Environment B was a workplace with eight computers. Two types of noises could be heard in this room. One is working computers' noise and the other of air conditioner. Glass-partition separated this computer class from the other computer class.

Environment C was a workplace with fifteenth computers. Also there were a blackboard, printer and air conditioner, that during experiment was turned off. Teacher's table stands in the middle of the room.

Environment D was a workplace with ten computers. Also there were a blackboard, printer and air conditioner, that during experiment was turned off. Teacher's table stands near the wall. There were no other noises in the room. Glass-partition separated this computer class from the other computer class.

Environment A

Environment B

Environment C

Environment D

Fig. 1. Environments used for measurement

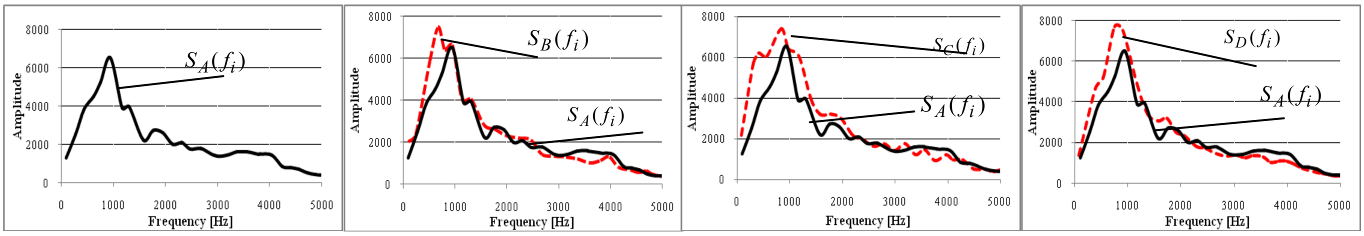

Fig. 2. Spectral characteristic $S(f_i)$ of environments A, B, C, D

Impulse signal $\Delta(t)$ was used to measure acoustic characteristics of the room. The impulse signal was created with a pistol shot. Acoustic response signal $y(t)$ to the signal $\Delta(t)$ was measured. Spectral characteristics $S(f_i)$ of signal $y(t)$ was computed:

$$S(f_i)=\frac{2\delta b^2}{(\sum_{j=0}^{p} a_j \cos 2\pi\delta j f_i)^2+(\sum_{j=0}^{p} a_j \sin 2\pi\delta j f_i)^2}, \qquad (1)$$

here $0 \le f_i \le \frac{1}{2\delta}$, $f_i = id\,(i=0,1,...,n)$, $d=\frac{F_{\max}}{n}$, $a_j(j=0,1,..,p), b$ – estimates of the coefficients of equation

$$y_t = \sum_{j=0}^{p} a_j y_{t-j} + bV_t \; (t=1,...,N) \qquad (2)$$

that describe digitized $y(t)$, when sampling interval is

$$\delta = \frac{1}{11025} s. \qquad (3)$$

Spectral characteristics $S(f_i)$ are presented in Fig. 2.

Experimental investigation

Speech command recognition systems used in the experiment were based on the dynamic time warping method [13]. Recognizer consists of three main parts: recording speech signal, training with etalons of speech command signal, and recognizing speech signal.

Recognizer uses dynamic programming for speech signal's endpoint detection. This method is based on the time warping and Bellmen function [14, 15].

Recognizer's training was based on the method of the nearest neighbor, when speech signal etalon with the lowest average distance to the other etalons is selected. Speech command signal etalon chooses the one who has a minimal distance to the other speech signal etalons.

To compute a distance between speech signals the following method was used: at first one speech signal's distance to other speech signals was computed. One signal was chosen as the basic one, and the others were left as comparative. This process was repeated with all speech signals. Finally, all candidate etalons for training were chosen. The second step was to group all etalons into groups of two, three, four, etc etalons. First step helped to find out which group is the best for etalon [13].

Dynamic time warping method was used to compute minimal distance between speech signals and etalons. Signal path was assessed using the Itakura rule. 80 speakers from different contingents aging from 20 to 55 took part in the experiment. Two speech signal classes $\Omega\{\text{Kaunas}\}$ next we will call it $\Omega\{K\}$ and $\Omega\{\text{Vilnius}\}$ next we will call it Ω $\{V\}$ were analyzed. Every speaker said about 25 speech signal examples. Analysis was made with Ω $\{K\}$ =2910 and Ω $\{V\}$ =2945 examples.. Signals were recorded in mono, 11025 Hz sampling rate, 16 bit amplitude quantization mode.

Recognition system's training

Two recognition situations were analyzed in the experiment. To train the class a set of 10 speech signal examples was used. Etalon signals created in the Environment A by woman voice at the same time, and different size sets of speech signal etalons were used for all training situations.

1st training situation. Two first speech signal etalons were taken from the set of speech signal for training. Recognizer used the nearest neighbor method to choose one the best signal for training. In this situation had a recognizer trained with one Ω class etalon of speech command.

2nd training situation. Six first speech signal etalons were taken from a set of speech signals for training. Recognizer used the same nearest neighbor method to choose five the best signals for training. In this situation we had a recognizer trained with five Ω class etalons of speech commands.

Experimental results

The experiment was performed with a set of speech signals Ω = {K, V}, four different environments, 80 speakers from different contingents and age, and three types of features:

- cepstrum coefficients;
- central cepstrum coefficients;
- coefficients of speech signal autoregressive process.

Two methods of training were analyzed. The first method, a class trained with one etalon of speech signal, and the second method, a class trained with five etalons of speech signal. Recognition was analyzed in two situations. In the first situation training and recognition are in the same environment. And in the second situation, training and recognition are taken in the different environments The results of this experiment are showed in Fig.3, Fig.4, and Fig.5.

The basic environment for creating etalons was Environment A (signed EA). One woman's voice was used to create speech signal etalons and to train the recognizer. Speech command recognition was made in the environments A, B, C and D.

After comparing the results in Fig.3, Fig.4 and Fig.5, it can be seen that the features of central cepstrum coefficient were the most neutral for the environment. In each environment situation recognizer gave better or very similar recognition results. Less neutral for environment features were coefficients of speech signal autoregressive process. It this situation a little bit worse recognition results were gotten, especially when used in different environments. Very sensitive features for environment were cepstrum coefficients. Using these features the worst recognition results were gotten, especially in situations with different environments.

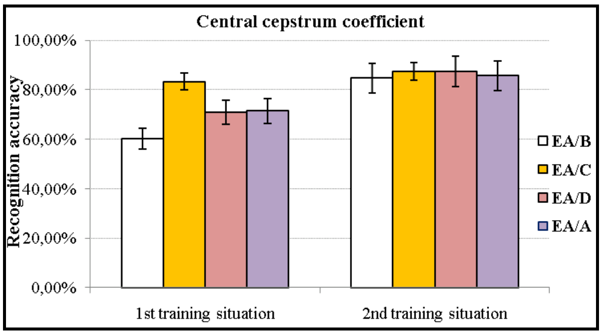

Fig. 3. Recognition results using central cepstrum coefficients features

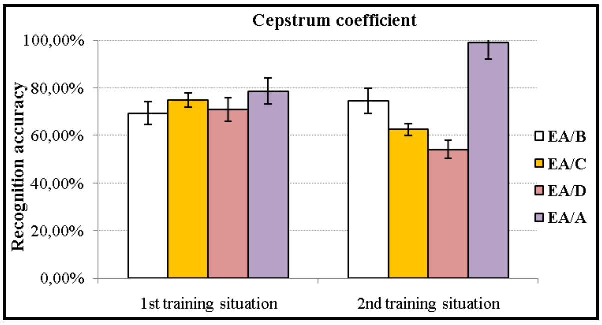

Fig. 4. Recognition results using cepstrum coefficients features

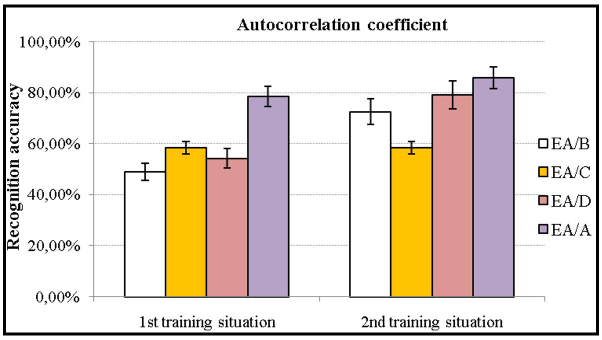

Fig. 5. Recognition results using coefficients of speech signal autoregressive process

The analysis of experimental results proved that the influence on speech signal recognition accuracy has speech signal etalons with specific feature used for training.

If recognition system is trained with a large set of speech signal etalons created in the same environment at the same time, it is possible that recognition accuracy will not improve

Conclusions

The analysis of the experimental results proved that:

1. The most environment neutral for speech signal training and recognition are central cepstrum coefficient features;
2. Very environment sensitive for speech signal training and recognition are cepstrum coefficient features;
3. On an average environment sensitive for speech signal training and recognition are coefficients of speech signal autoregressive process.

Recognition accuracy will be improved if environment for recognition and for training is the same, if a class of etalons is trained with a quite large set of speech signals (the recommended size of the etalon set would be about five or more etalons), and if it is applied to all features: central cepstrum coefficient, cepstrum coefficient and coefficients of speech signal autoregressive process.

Recognition accuracy will not always improve if different environments for recognition and for training are used, and if the largest set of speech signals is used to train speech command.

References

1. **Schötz S.** Linguistic & Paralinguistic Phonetic Variation in Speaker Recognition & Text–to–Speech Synthesis // Term paper for course in Speech Technology, GSLT, 2001.
2. **Denda Y., Nishiura T., Kawahara H.** Noisy speech recognition with microphone array steering and fourier/wavelet spectral subtraction // IEEE, 2003.
3. **Das S., Bakis R., NBdas A., Nahamoo D., Picheny M.** Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system // IEEE, 1993.
4. **Stahl V., Fischer A., Bippus R.** Acoustic synthesis of training data for speech recognition in living room environments // Acoustics, Speech, and Signal Processing, IEEE International Conference on, 2001. – Vol. 1. – P. 285–288.
5. **McCowan I. A., Marro C., Mauuary** L. Robust speech recognition using near field superdirective beamforming with post filtering // IEEE, 2000.
6. **Darren C., Iain M., McCowan A., Moll**e **D.** Microphone array speech recognition: experiment on overlapping speech in meeting // IEEE, 2003.
7. **Furui S., Iwano K., Hori C., Shinozaki T., Saito Y., Tamura S.** Ubiquitous speech processing // IEEE, 2001.
8. **Shukla S., Mahadeva Prasanna S. R.** Speech Recognition under Stress Condition // IIT, Guwahati, 2009.
9. **Fink G. A., Hohenner S.** Experiments in Distant Talking Speech Recognition Using a Standard Database. – 2005.
10. **Baker J., Deng Li, Glass J., Khudanpur S., Chin–hui Lee, Morgan N., Oapos Shaughnessy D.** Developments and directions in speech recognition and understanding // Signal Processing Magazine, IEEE, 2009. – Vol. 26. – Iss. 3. – P. 75–80
11. **Baker J., Deng Li, Glass J., Khudanpur S., Chin–hui Lee, Morgan N., Oapos Shaughnessy D.** Updated MINDS report on speech recognition and understanding // Signal Processing Magazine, IEEE, 2009. – Vol. 26. – Iss. 4. – P. 78–85
12. **Furui S.** Recent advances in spontaneous speech recognition and understanding // Proc. SSPR2003. – Tokyo, Japan, 2003. – P. 1–6.
13. **Paulikas Š., Karpavičius R.** Application of Linear Prediction Coefficients Interpolation in Speech Signal Coding // Electronics and Electrical Engineering. – Kaunas: Technologija, 2007. – No. 8(80). – P. 39–42.
14. **Tamuliavičius G., Lipeika A.** Žodžio pradžios ir galo nustatymas atpažįstant atskirai sakomus žodžius // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2004. – Nr. 2(58). – P. 61–64
15. **Lipeika A, Lipeikienė J.** Word Endpoint Detection Using Dynamic Programming // Informatica, 2003. – Vol. 14. – No. 4. – P. 487–496.

Received 2010 02 15

G. Čeidaitė. L. Telksnys. Analysis of Factors Influencing Accuracy of Speech Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 9(105). – P. 69–72.

Factors influencing accuracy of speech recognitions is investigated. The main attention was given to environment, training conditions and features . The results of the influence of the factors to the accuracy of speech recognition are presented. The analysis of experimental results proved that the biggest influence on recognition accuracy has environments' in which speech commands' recognition are used and size set of etalons of speech commands used for training. Ill. 5, bibl. 15 (in English; abstracts in English and Lithuanian).

G. Čeidaitė. L. Telksnys. Atpažinimo tikslumui įtakos turinčių veiksnių analizė // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2010. – Nr. 9(105). – P. 69–72.

Nagrinėjami šnekos signalų atpažinimo tikslumui įtakos turintys paralingvistiniai veiksniai. Analizuojamos patalpų etalonų sudarymo sąlygos ir trijų požymių pritaikymo įtaką. Pateikiami atlikto tyrimo rezultatai, rodantys įvairių veiksnių įtakojimą atpažinimo tikslumui. Nustatyta, kad didžiausią įtaką šnekos signalų atpažinimo tikslumui turi aplinka, kurioje atpažįstami šnekos signalai, ir komandų etalonams sudaryti vartojamų žodžių skaičius. Il. 5, bibl. 15 (anglų kalba; santraukos anglų ir lietuvių k.).