Use of voice identification technology for speaker segmentation of communications between a controller and a pilot

Grigory Tyschenko
GritTec Laboratory (GritTec Ltd.)
08/2016

ABSTRACT
This article describes method of automatic speaker segmentation in the audio recordings of communications between a controller and a pilot. The method is based on the automatic voice identification technology of target controller in offline mode.

Keywords
Speaker Segmentation, Voice Identification, K-means, Gaussian Mixture Models, GMM, Mel Frequency Cepstral Coefficients, MFCC, False Acceptance Rate, FAR.

1. INTRODUCTION
In recent years there is growing interest in methods of automatic voice segmentation of communications between a controller and pilots. This interest is due to the increased safety level of air flights and to the control by monitoring services.

Conventionally, all the speaker methods can be divided into 2 types. The first type includes all methods used when the quantity of speakers is known in advance. These methods are based on the selection of acoustic features, with their further clustering, for example, K-means, Gaussian mixture models (GMM) [1]. Application of these methods involves the use of various estimations and decision criteria for correlation of the analyzed time interval with the target speaker, for example, Bayesian Information Criterion (BIC) [2].

The second type includes methods used when the quantity of speakers is not known in advance. Among them there are methods that track the dynasmic changing of the acoustic features on the base of autoassociative neural networks (AANN) and methods of voice identification [3, 4, 5].

This article describes a method of speaker segmentation on the base of automatic voice identification of target speakers (controllers) in offline mode.

2. SPEAKER SEGMENTATION
The target voice models of target speakers (controllers) were loaded in system for identification. The voice model for each target speaker (controller) consists of:

- GMM model;
- pattern of error model with FAR estimations [6];
- and model of background noises and channel distortion which is used for normalization of acoustic features vector [7, 8].

The stream of speech signal in the format (PCM 8kHz, 16bits linear Mono) is devided in frames with duration 20 ms and 10 ms overlap. Mel Frequency Cepstral Coefficients (MFCC) , their 1-st and 2-st derivatives are calculates for each vocal frame:

$$x[j] = (MFCC+\Delta FCC+\Delta\Delta MFCC)$$

The calculated vector of features for analyzed j-th frame is compared with the target voice models of controllers a[k] and in result probavility of observation P(x[j]|a[k]) of current target controller a[k] is calculated. Then the value of probability of observation is projected on the sample of error model FAR of current a[k] controller:

$$FAR[j, k] = F(P(x[j]|a[k])),$$

In result it is determined the value of acceptance error for current target controller a[k] on the j-th frame. The error value is compared with a threshold error value (thr_far ~ 5%), on the basis of what it is determined the correlation of the current frame with the target voice controller or the voice of a pilot. Then the identified frames with step duration of 0.5 seconds are averaged over the status of acceptance to the target voice controller a[k], in result it is determined the presence of the target voice controller or voice of unknown pilot.

3. EXPERIENCE RESULTS
The proposed speaker segmentation method was checked on actual audio recordings of communications between controllers and pilots for English language. Example of speaker segmentation based on FAR estimation for target controller is shown on Fig.1.

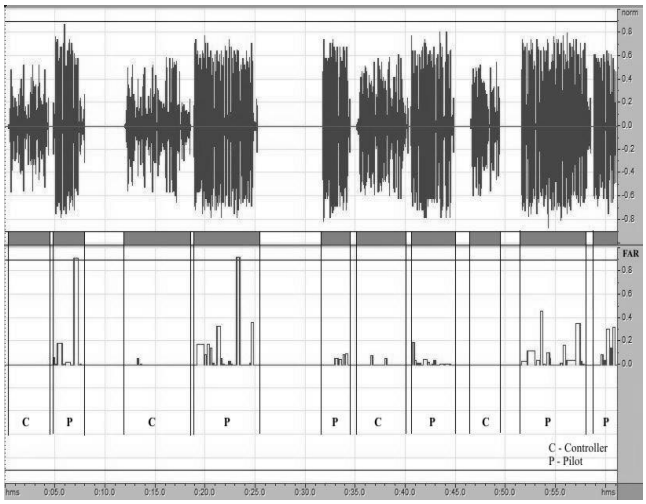

Fig. 1. Speaker segmentation on the base of FAR estimation.

4. CONCLUSION
The proposed speaker segmentation algorithm has shown good results during testing of the real audio recordings of communications between target controllers and pilots. Further it is planned to automate the process of determining of the FAR threshold (thr_far ~ 5%) of correlation of the analyzed frame with the target controller.

5. REFERENCES

[1] L. Lu, H. J. Zhang. 2002. “Speaker change detection and tracking in real time news broadcasting analysis”. In Proceedings of the 10th ACM Int’l conf. Multimedia. 602–610.

[2] S. S. Chen, P. S. Gopalakrishnan. 1998. “Speaker, environment and channel change detection and clustering via the Bayesian information criterion”. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

[3] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman. “Speaker identification using mel frequency cepstral coefficients”. 3rd International Conference on Electrical & Computer Engineering

[4] S. Jothilakshmi S. Palanivel V. Ramalingam. "Unsupervised Speaker Segmentation using Autoassociative Neural Network". 2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7.

[5] Douglas A. Reynolds, Richard C. Rose. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 3, NO. , JANUARY 1995.

[6] GritTec's Speaker-ID: Automatic Text Independent Speaker Identification. DOI= http://www.grittec.com/speaker-identification.html

[7] M.J.F. Gales & S.J. Young. "Robust continuous speech recognition using parallel model combination". EDICS Number: SA 1.6.8, Cambridge Unversity Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England.

[8] Svein G. Pettersen, Magne H. Johnsen, Tor A.Myrvoll. "Joint Bayesian Predictive Classification and Parallel Model Combination for Robust Speech Recognition". Department of Electronics and Telecommunications Norwegian University of Science and Technology.