Speaker identification for aeronautical communications

Sara SEKKATE
LIM@II FSTM
sarasekkate@gmail.com

Mohammed KHALIL
LIM@II FSTM
medkhalil87@gmail.com

Abdellah ADIB
LIM@II FSTM
adib@fstm.ac.ma

Abstract

In this paper, we aim to fulfill this goal by developing a Speaker Identification System (SIS) for future aeronautical communications systems. Furthermore, we present a novel feature extraction scheme based on multi-resolution analysis. The proposed system, called i-SMFCC uses Mel Frequency Cepstral Coeficients (MFCC) features of Stationary Wavelet Transform (SWT) sub-bands. The extracted features are modeled using the i-vector approach and Support Vector Machines (SVM) are adopted as a back-end classifier. The performance of the proposed SIS is evaluated using two publicly available databases. Comparison of the proposed approach with the baseline MFCC feature extraction shows the feasibility and the robustness of the proposed method.

1 Introduction

Speaker recognition is a well-established research problem and has found use in many applications, including voice authenticated bank transaction, access control, and prison call monitoring [1]. However, research has been rarely devoted to the integration of Automatic Speaker Recognition (ASR) in the aeronautical industry [2].

In Air Trafic Control (ATC), voice communication serves as the main media for delivering instructions and important information between pilots and controllers [3]. The road to improving safety in ATC, therefore, definitely passes through improving the air-ground communications safety. By far, the most prominent issue surrounding these communications is the heightened risk of callsign confusion [4]. As a consequence, it is possible for the pilot to accept clearances meant for others, leading to wrong subsequent actions and incidents with a high potential to cause death [5].

In this paper, we aim to develop a closed-set[1] text-independent[2] speaker identification[3] for the En-route[4] airspace that could be used to prevent call-sign confusion and hence increase flight safety.

2 The proposed approach

The proposed system is summarized as following:

[1]Closed-set systems suppose that the unknown speaker to be identified is known a priori to be one of the registered speakers set.

[2]Text-independent mode imposed no restrictions on spoken phrases, and hence the speaker can utter any word in order to be recognized.

[3]The goal is to identify an input speech by selecting one model from the previously enrolled speaker models.

[4]En-route scenario describes the state where the aircraft is airborne and communicates with the control tower (air-ground communication) or with an other airplane (air-air communication).

1. First, for each speech sample, the signal is decomposed into 9 sub-bands using SWT[5] to achieve a series of approximation and detail coefficients.
2. Then, MFCC are computed from each sub-band [6]. Here, the first 13 coefficients derived from a 20-channel mel-scaled filterbank are extracted from speech frames of 25ms with a frame shift of 10 ms, removing the first one because it carries less speaker specific information. In addition to static MFCC features, the log-energy as well as first and second derivatives were also included to produce a feature vector of 39 elements.
3. Concatenate all the sub-band features to produce a final feature vector, denoted as SMFCC.
4. Repeat step 1 to step 3 for each speech sample to create a feature matrix that will be fed into the i-vector modeling framework [7].
5. The resulting feature vectors are then assembled, modeled using the i-vector approach and fed to the SVM classifier.

3 Experimental results

The performances of the proposed system are evaluated on two publicly available datasets. The first one consists of 10 speakers (4 females and 6 males) from ATCOSIM speech corpus [8]. The second one is a set of 455 speakers (41 females and 414 males) from Voxforge database [9].

From the results listed in Table 1 for ATCOSIM database, we observe an increase of 5 to 11 percent in the accuarcy at 5dB with the highest accuracy of 49% reached for SMFCC features without i-vector modeling. We can also see that in general, the performance of i-MFCC is lower than MFCC at all SNR ranges, while i-SMFCC outperforms SMFCC beyond 10dB. On the other hand, Table 2 reports the results obtained using Voxforge database, where the advantage of using SMFCC is again obvious compared to baseline MFCC. Except for 5dB, although there is an improvement compared to MFCC, the achieved performance remains poor.

Table 1: Accuracy (%) of the proposed method in noisy environments using ATCOSIM speech corpus

Transmission Channel	**SNR (dB)**	**ATCOSIM**			
		MFCC	**SMFCC**	**i-MFCC**	**i-SMFCC**
AWGN	5	45.67	**49**	14.67	25
	10	84.33	92.67	81.67	**96.33**
	15	84.33	92.67	81.67	**96.33**
En-route	Infinite	84.33	92.67	81.67	**96.33**
En-route + AWGN	5	**37**	36	10.33	15.33
	10	84	89	80.33	**93**
	15	84.67	93	81.67	**96.33**

Table 2: Accuracy (%) of the proposed method in noisy environments using Voxforge speech corpus

Transmission Channel	**SNR (dB)**	**Voxforge**			
		MFCC	**SMFCC**	**i-MFCC**	**i-SMFCC**
AWGN	5	2.12	**10.84**	3	5.71
	10	24.47	71.79	79.63	**92.45**
	15	24.47	71.87	79.63	**92.45**
En-route	Infinite	24.47	71.87	79.56	**92.45**
En-route + AWGN	5	1.39	**8.64**	1.47	2.56
	10	22.12	66.81	75.53	**84.40**
	15	24.40	71.72	79.56	**92.45**

References

[1] Zhao, X. & Wang, Y. & Wang, D. (2014): Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(4), 836-845*.

[5]The main reason of using SWT is the fact that it is time-invariant transform as compared to DWT.

[2] Hofbauer, K. & Hering, H. & Kubin, G. (2005): Speech watermarking for the VHF radio channel. *4th EUROCONTROL Innovative Research Workshop*

[3] Haas, E. (2002) Aeronautical channel modeling. *IEEE Trans. on Vehicular Technology 51(2), 254-264*

[4] Delpech, E. & Laignelet, M. & Pimm, C. & Raynal, C. & Trzos, M. & Arnold, A. & Pronto, D. (2018) A Real-life, French-accented Corpus of Air Traffic Control Communications. *Language Resources and Evaluation Conference (LREC), Miyazaki, Japan*

[5] Chen, S. & Kopald, H. & Chong, R. & Wei, Y. & Levonian, Z. (2017) Read Back Error Detection using Automatic Speech Recognition *Twelfth USA/Europe Air Traffic Management Research and Development Seminar*

[6] Nagaraja, B.G. & Jayanna, H.S. (2012) Multilingual Speaker Identification with the Constraint of Limited Data Using Multitaper MFCC. *Recent Trends in Computer Networks and Distributed Systems Security pp. 127-134*

[7] Kheder, W.B. & Matrouf, D. & Bousquet, P.M. & Bonastre, J.F. & Ajili, M. (2017) Fast i-vector denoising using map estimation and a noise distributions database for robust speaker recognition. *Computer Speech and Language pp.104-122*.

[8] Hofbauer, K. & Petrik, S. & Hering, H. (2008) The ATCOSIM corpus of non-prompted clean air trafic control speech. *Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech, Morocco*

[9] Voxforge database. Tech. rep. URL http://voxforge.org