Research of Speech Signal Acoustic Models for Speaker Recognition

Greta Borcovaite

Siauliai University,
Faculty of technology,
Physical and biomedical sciences

Abstract— **The research of speech signal acoustic models for speaker recognition been described in this paper. The aim of this research – to investigate acoustic speech signal models suitable for speaker recognition. In the analytical practical part, voice records were investigated, MFCC features were extracted, acoustic speech signal models were trained and tested.**

Furthermore, investigation results have shown that components in records distributed differently. The six most common acoustic models components were chosen. The most common voice and background components are different. Statistical analysis has shown that log-likelihoods are not statistically significant different for different languages records when same type and same languages acoustic models were applied. Moreover, log-likelihoods are not statistically significant different for different languages records when English acoustic models were used. Finally, log-likelihoods differ mostly in Spanish and English language records. Increasing the number of English and Spanish records log-likelihoods are statistically significant different when English acoustic models are used.

Index Terms— **MFCC features, GMM, speaker recognition, acoustic model.**

I. INTRODUCTION

Biometrics – measurement and statistical analysis of people's personal physical and behavioral characteristics. Types of biometrics [1]: DNA matching, ear, eye – iris recognition, eye – retina recognition, face recognition, fingerprint recognition, fingerprint geometry recognition, gait, hand geometry recognition, odor, signature recognition, typing recognition, vein recognition, voice or speaker recognition. Voice biometric uses personal unique speech characteristics such as physiology of the vocal tract and specific speech features, their comparison to determine the number of speakers, language, speaker recognition, dialect or accent determination.

Speech signal acoustic models are used for speaker recognition. Acoustic models are statistical models that classify records to classes of models. Models are trained using different age and gender audio records. The length of training audio records usually takes several hours.

HTK [9] and Kaldi [2] are widely used for various voice technology tasks. These open source packages could be applied for audio signal analysis and phonetic alignment i.e. for phoneme boundaries segmentation in given transcription. MSR Identity Toolbox [6] can be used for speaker recognition.

Freely available acoustic model [9] can be used for speaker recognition. The model has 2048 classes of 60 features. However, for which languages model fit is not clear.

The aim of this paper was to investigate acoustic speech signal models suitable for speaker recognition. Methods of acoustic models creation and results of various languages records described below.

II. METHOD

Acoustic model depends on selected features. Mel-Frequency Cepstral Coefficients (MFCC) are widely used in up-to-date speaker recognition systems [9].

Feature extraction is one of the most important steps in speaker recognition. Thus, feature extraction steps: Mel filter bank calculation, MFCC features determination, first and second order differential features addition to obtained features, voice activity detection, MFCC feature normalization using averages and standard deviations of MFCC features from all audio recordings.

Using normalized features GMM likelihoods were determined, GMM maximum likelihoods were chosen, voice and background components were separated, and voice features were detached. Moreover, using voice features acoustic models were created.

MFCC features and sampling frequency of 8 kHz were used in research.

Acoustic models are using Gaussian Mixture Models (GMM) [3]. Baum-Welch algorithm can be used for model parameter estimation [4]. GMM can be applied for automatic speech recognition. Feature probabilities can be evaluated using this method.

Feature extraction method was taken the same as in acoustic model [8].

III. CREATING ACOUSTIC MODELS

Before you begin to format your paper, first write and save Expectation – maximization (EM) algorithm is a method to find maximum likelihood estimates of parameters in GMM [5]. The algorithm is an iterative algorithm that starts from some initial estimate and then proceeds to the number of iterations until the difference between adjacent values becomes sufficiently small.

The algorithm accomplished in 12 model adjustment and incremental iterations. Once all iterations were completed, acoustic speech model was returned.

In addition, k-means clustering method was chosen for acoustic model creation [7]. Training performed in 21 iterations.

IV. RESEARCH RESULTS

Polylingual speech acoustic model for speaker recognition was created during the research. English GMM model was applied for speech feature extraction. Extracted features was used to build GMM and k-means acoustic models in English, Spanish, Italian, French, Russian and German languages (see Table 1).

Table 1 Acoustic models

Denotation	Acoustic models
1	English GMM model
2	English k-means model
3	Spanish GMM model
4	Spanish k-means model
5	Italian GMM model
6	Italian k-means model
7	French GMM model
8	French k-means model
9	Russian GMM model
10	Russian k-means model
11	German GMM model
12	German k-means model

Repetition rates in audio records for speech acoustic models components were determined. The investigation of most frequently repeated various acoustic models components was done. As a result was found that components in records distributed unevenly. Assessing models, six most common acoustic models components were chosen. Ascertained that six most commonly repeated English GMM model components occupies 12.5% of voice areas in female records and 9% – in male records in English. Five of them are common. The most commonly repeated component also common. The same position it takes in male records in French and in female records in Russian and German languages. For other investigated languages, it falls into most commonly repeated component six. In addition, there is no such pattern for k-means method.

Voice and background components analysis shown that voice and background components are not the same. Thus, voice components match the speech.

Log-likelihoods were estimated using various acoustic models for each considered language. Therefore, 10 testing records for each language were chosen.

One-way analysis of variance (one-way ANOVA) was used to test if various acoustic models log-likelihoods are statistically significant different for various languages with a significance level of 0.05. Tested, if log-likelihoods are statistically significant different for different languages records when 1st acoustic model was used. Since p-value 0.045 is less than 0.05, log-likelihoods are statistically significant different for different languages records when 1st acoustic model was used with a significance level of 0.05. Also noticed that log-likelihoods of Spanish language records differ from other languages results. For this reason, one-way ANOVA test was repeated without this language. Seeing that p-value 0.08 is more than 0.05 was obtained that log-likelihoods are not statistically significant different for different languages records when 1st acoustic model was used with a significance level of 0.05.

Summarizing log-likelihoods results of different languages audio records using 2nd acoustic model was tested if log-likelihoods are statistically significant different for different languages records when 2nd acoustic model was used. Obtained that p-value 0.083 is more than 0.05. Thus, log-likelihoods are not statistically significant different for different languages records when 2nd acoustic model was used with a significance level of 0.05.

GMM acoustic speech models in Spanish, Italian, French, Russian and German languages was built using voice features of those languages. Log-likelihoods of various languages records were estimated using GMM models in chosen languages. Statistical analysis performed using obtained results. Tested, if log-likelihoods are statistically significant different for different languages records when GMM acoustic models in these languages were used. Obtained that p-value 0.061 is more than 0.05. Thus, log-likelihoods are not statistically significant different for different languages records when GMM acoustic models in these languages were applied with a significance level of 0.05.

Spanish, Italian, French, Russian and German languages k-means acoustic models were created using voice features of those languages. Log-likelihoods of different languages records estimated using k-means acoustic models in chosen languages. Using achieved results one-way ANOVA test was performed. Checked, if log-likelihoods are statistically significant different for different languages records when k-means acoustic models in these languages were applied. Hence, p-value 0.057 is more than 0.05, log-likelihoods are not statistically significant different for different languages records when k-means acoustic models in these languages were applied with a significance level of 0.05.

One-way ANOVA test shown that log-likelihoods are not statistically significant different for different languages records when same type and same languages acoustic models were used.

Student's t test was used to test if log-likelihoods are statistically significant different for compared languages when English acoustic models were used with a significance level of 0.05. Thus, English and other testing language (Spanish, Italian, French, Russia

nor German) was compared. Tested, if log-likelihoods are statistically significant different for various languages records when English acoustic models were used (see Table 2).

Table 2 Student t test results

Compared languages	p-value	
	Model 1	Model 2
English/ Spanish	0,118	0,108
English / Italian	0,235	0,230
English / French	0,151	1,000
English / Russian	0,627	0,623
English / German	0,227	0,291

Obtained, that log-likelihoods are not statistically significant different for different languages records when English acoustic models were used with a significance level of 0.05.

Moreover, English acoustic models averages of log-likelihoods analysis was made. Checked which language differ the most from English. For this reason, averages of log-likelihoods in different languages were estimated when acoustic models in English were used.

Using estimated averages of log-likelihoods obtained that log-likelihoods differ mostly in Spanish and English. Therefore, 100 additional testing records in English and Spanish were selected and their log-likelihoods were estimated using English acoustic models. Student t test performed using obtained results where compared languages – English and Spanish, significance level – 0.05. Tested if increasing the number of English and Spanish records log-likelihoods *are* statistically significant different *when* English acoustic models were used. Obtained, that p-value of 1st model and 2nd model (both are equal to 0.000) is less than 0.05. Thus, increasing the number of English and Spanish records log-likelihoods are statistically significant different when English acoustic models were used with a significance level of 0.05.

V. CONCLUSIONS

1. Components in records distributed unevenly. Assessing models six most commonly repeated components selected. Ascertained that six most commonly repeated English GMM model components occupies 12.5% of voice areas in female records and 9% – in male English records. Five of them are common. The mostly repeated component also common. The same position it takes in male French records, female Russian and German records. For other investigated languages, it falls into most commonly repeated component six;
2. Most commonly repeated voice and background components are different;
3. One-way ANOVA test shown that log-likelihoods of different languages audio records have no statistically significant difference from a language applying same language and same type acoustic models;
4. As a result of English acoustic models log-likelihoods means analysis was obtained, that log-likelihoods differ mostly in Spanish and English records;
5. Student t test shown that log-likelihoods are not statistically significant different for different languages records when English acoustic models were used. Increasing the number of English and Spanish records log-likelihoods are statistically significant different *when* English acoustic models were used.

REFERENCES

[1] T. K. Ghazali, N. H. Zakaria, "Security, Comfort, Healthcare, and Energy Saving: A Review on Biometric Factors for Smart Home Environment," in Journal of Computers, 2018, vol. 29, no. 1, pp. 189-208.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burger and other, "The Kaldi Speech Recognition Toolkit," IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

[3] L. R. Rabiner, B. H. Juang, Fundamentals of Speech Recognition. Upper Saddie River, NJ: Prentice-Hall, 1993, p. 507.

[4] G. Raškinis, D. Raškinienė, "Development of Medium-vocabulary Isolated-Word Lithuanian HMM Speech Recognition System," Lithuania, Vilnius, Informatica, 2003, vol. 14, no. 1, pp. 75-84.

[5] T. Ruzgys, "Daugiamačio pasiskirstymo tankio neparametrinis įvertinimas naudojant stebėjimų klasterizavimą," PhD thesis, Lithuania, Vilnius, Institute of Mathematics and Informatics, 2007, p. 94.

[6] S. O. Sadjadi, M. Slaney, L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," IEEE SLTC Newsletter, 2013.

[7] R. Salman, "Contributions to k-means Clustering and Regression via Classification Algorithms," PhD thesis, Virginia Commonwealth University, Department of Computer Science, 2012, p. 98.

[8] Voice Biometry Standardization Initiative [interactive] [last viewed 2018-05-23]. Internet access: <http://www.voicebiometry.org>.

[9] S. J. Young, P. C. Woodland, W. J. Byrne, The HTK Book. UK: Cambridge University Engineering Department, 2015, p. 355.