

Information Visualization, Exploration, and Error Analysis in the NIST 2010 Speaker Recognition Evaluation

Xinhui Zhou¹, Daniel Garcia-Romero², Shihab Shamma³, Carol Espy-Wilson⁴

Department of Electrical and Computer Engineering, University of Maryland, College Park, USA

¹zxinhui@umd.edu, ²dgromero@umd.edu, ³sas@umd.edu, ⁴espy@umd.edu

Abstract—In the NIST speaker recognition evaluation, a lot of research has been done on optimizing the front-end acoustic features and the back-end classifiers in terms of equal error rate (EER) and detection cost function (DCF). However, there is no much further analysis on the evaluation database and on how the error patterns are correlated to various speaker and channel properties, and how the error patterns differ among different acoustic parameters and different classifiers. In this preliminary study, we took advantage of an information visualization and exploration tool (TIBCO Spotfire) to facilitate such analyses. All the key information for speaker, segment, and trial in the NIST SRE 2010 extended-core task was integrated together, along with the scores from our probabilistic linear discriminant analysis (PLDA) system. Thus we are able to conveniently visualize and exploit various types of data in many ways. Furthermore, the corresponding segments and voice activity detection (VAD) results are automatically linked for inspection and comparison. Some irregular or mislabelled data was found through data visualization. Our preliminary analysis on conditions 2 and 5 shows that the miss detection errors are only from a small portion (about 20% or less) of the speakers in trials, whereas most of the speakers are involved in false alarms. The average non-target trial score decreases monotonically with the age difference between the two speakers, and most of the false alarms (about 80%) are involved in those trials where the age difference is less than ten years old. We believe that further work on error analysis, facilitated by visualization and exploration tools, will provide us insights in understanding our speaker recognition system performance, designing the evaluation database, and potentially improving the system performance¹.

Keywords—Speaker recognition, NIST SRE10, information visualization, Spotfire, error analysis, PLDA.

I. INTRODUCTION

In the NIST speaker recognition evaluation, a lot of research has been done on optimizing front-end acoustic features and back-end classifiers in terms of equal error rate (EER) or detection cost function (DCF) [1][2][3][4]. However, there is no much further analysis (except [5][6]) on the

properties of speaker population and their speech data in the evaluation database, on how those miss detection and false alarm errors are correlated to various speaker and channel properties, and how the error patterns may differ among various acoustic parameters and various classifiers.

In this preliminary study, we took advantage of an information visualization and exploration tool TIBCO Spotfire [7] to facilitate such analyses. All the key information provided by NIST for speakers, segments, and trials in the NIST SRE 2010 extended-core task was integrated, along with the scores from our speaker recognition system. Through analyzing the key information and the trial scores produced by the speaker recognition systems together, we tried to gain insights in understanding the evaluation database and the speaker recognition systems. We hope, in the long run, those insights will help us bridge the gap between the dominant data-driven approaches in speaker recognition and the scientific knowledge in speech science about speaker characteristics [8][9], and eventually improve our speaker recognition performance.

In the rest of this paper, we describe our experiment setup and acoustic features in the NIST SRE10 extended core task [10], how the key information was used, and the main features of the data visualization tool Spotfire. Then we present our preliminary results on analysing the speaker population, the speech data properties in terms of speech type and channel type, the scores, and the errors. Finally, a summary is given.

II. EXPERIMENT SETUP, KEYS AND VISUALIZATION TOOL

A. Experimental Setup

Our speaker recognition system is an I-vector-based gaussianized probabilistic linear discriminant analysis (PLDA) system [11]: both the i-vector extractor and the PLDA system were gender-dependent. Sufficient statistics was collected using the same 2048 mixture UBMs. The subspace matrix T with 400 columns was obtained by pooling together all the telephone and microphone recordings in the development set (see [11] or [12] for details) from the corresponding gender. For the PLDA model, the same data was used (excluding the Fisher database) to train the eigenvoice matrix Φ with 200 columns and the full-covariance matrix $\Sigma \in \mathbb{R}^{(400 \times 400)}$.

¹This research was partially funded by NSF award #0917104 and by the Office of the Director of National Intelligence (ODNI) and Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory (ARL).

The acoustic parameters we used in this study are linear frequency cepstral coefficient (LFCC) [12] and mel-frequency cepstral coefficient (MFCC). For LFCC, the speech signal is band-limited to 300-3400 Hz and 24 filter-banks were used. The 19 cepstral coefficients plus its delta make a 38-dimensional feature vector. The analysis window is 20 ms with a shift of 10 ms. The cepstral mean subtraction and variance normalization was applied. For MFCC, the 19 cepstral coefficients and frame energy plus its delta and double delta make a 60-dimensional feature vector. A short-time gaussianization was applied. The voice activity detection (VAD) for segmenting speech from the silence region is based on the ASR transcript combined with the output of an energy-based VAD system.

The DET curves of our PLDA system on the NIST SRE10 extended-core task are shown in Fig. 1 a) and b).

For the rest of the paper, the scores and EERs used for analysis are from MFCC unless LFCC is mentioned.

B. Key information in the database

In the NIST SRE10 extended-core task, information provided about the trials is limited to speaker gender, speech type (interview/phonecall), basic channel type (microphone or telephone line). After the evaluation, detailed key information about speakers, segments and trials were provided by NIST. Information for each speaker includes age, height, weight, native-language, state, dialect and others. Information for each segment includes speaker identity, channel type (microphone type and its index in the recording room, telephone type, phone number id) and other additional channel properties such as signal quality.

In this study, three key tables were created for speakers, segments and trials, respectively. The speaker table has fields for various speaker properties and each row is for one speaker. The segmental table has fields for the segmental properties and also for the corresponding speaker's information. The trial table includes properties of the two segments in each trial and also the scores produced by the PLDA system. Audio files and their corresponding VAD files are linked through hyperlinks in the tables.

C. Visualization Tool

Spotfire [7], a popular and powerful data visualization and exploration tool, was used to analyse these three key tables. Its user-friendly interface (shown as in Fig. 2) enables us to analyse data without any effort in development. It takes tables as its input and users can conveniently visualize and explore the data in various ways such as box plot, scatter plot, bar plot, graph in terms of original data or statistical results. Data visualization helps us find normal patterns and spot outliers in data. We can easily filter out certain data and only focus on particular subset we are interested in. In particular, the audio files can be accessed directly by clicking the hyperlinks in the table, which allows us to analyse specific audio files.

The results of analysing the three tables are presented in this section. All the figures are produced by Spotfire.

A. Speaker

Fig. 3 shows the speaker distribution in gender and age. There are total 446 speakers in the NIST SRE10 database, 210 males and 236 females. 433 are native English speakers, and 410 are born in USA. Among them, 226 are raised in Pennsylvania, 36 in New York, and 35 in New Jersey. There are 73 smokers. Most speakers are 20 to 35 years old. The mode age of male is 24 and female 26. The oldest speaker is 92 years old and the youngest is 19.

As we can see, most of the speaker population in the SRE10 database are at young age and most of them are raised in those three states. This should be kept in mind when we compare speaker recognition performances from different databases. In addition, the speaker properties such as age, height, weight, smoker, dialect should be factors to consider in error analysis.

B. Segment

Fig. 4 a) b) and c) show the number of segments in terms of speech type, channel type, and speaker id. Fig. 4 d) shows how many speakers in each channel. Out of 11 microphones in the recording room [13], data in Mic02, 04, 05, 07, 08, 12 and 13 were used in SRE10. Mic04 and Mic08 are also used for recording some telephone call conversations. Unknown channel 'tel' only appears in test segment and there is no speaker information available for those files. There are much more data from cell phones than from landline telephones. Only 100 speakers used cordless landline telephone, whereas almost all of speakers used cell and corded landline telephone. Most speakers have 14-21 interview segments and 14-22 phonecall segments.

Fig. 5 shows the number of speakers using each phone ID and the number of phone id each speaker used. It can be seen that there are 5 phone ids most frequently used by the speakers (2 cell phones and 3 corded-landline telephones). Most speakers used 2-3 different phones for their telephone calls.

Fig. 6 shows the number of segments in high and low vocal effort. Segments in high and low vocal effort only exist in corded land-line phone with three telephone IDs in total.

Fig. 7 shows the histograms of speech frame numbers in both interview and phonecall. There are two modes in interview data for 3 min and 8 min long segments respectively. It can be seen that there are a number of files with zero or very few frames. Most of those files just contain noise or pure tone. In average, 40% of frames in phonecall data are detected as speech.

C. Trial

Fig. 8 shows the number of target and nontarget trials for each of the nine conditions. It can be seen that there are many more nontarget trials than target trials, and the number of trials varies a lot across conditions. In determining the required number of trials, there is a rule called 'the rule of 30'[5]: to be 90% confident that the true error rate is within 30% of the observed error rate, there must be at least 30 errors.

Based on this rule and the EERs shown in Fig. 1a), the required number of target trials should be: 1898, 1160, 1195, 1694, 1638, 717, 539, 1707, 2736 for each of the nine conditions respectively. The actual numbers of target trials in condition 7 and 9 are fewer than the required ones. If we would like to have a more strict rule like 'be 90% confident that the true error rate is within 10% of the observed error rate', the required number of target trials should be: 17270, 10554, 10877, 15413, 14905, 6525, 4907, 15532, 24897. In this case, only condition 2 has the required number of target trials. So due to the limited trial numbers, the results from the speaker recognition systems should be taken and interpreted with caution.

Fig. 9 shows the number of trials for each speaker in models, and the unique models per speaker. To check if each speaker has any error pattern, each speaker should have adequate number of independent models and trials.

D. Score of trials

Fig. 10 shows the trial score histograms for different cases. It can be seen that the histograms for target trials are not smooth due to the limited number of trials and the histogram shape for nontarget trials is asymmetrical.

Fig. 12 shows the trial scatter plots for speakers in model. It can be seen that very few target trials and speakers are involved in miss detection, whereas most speakers are involved in false alarm. Each trial or trials for one specific speaker can be further checked and analysed through the scatter plots. For example, we found that the lowest score in target trials for male speaker in condition 5 is caused by the cross-talking in the telephone channel, and the VAD includes the other speaker's speech.

Fig. 11 shows the score scatter plot for both LFCC and MFCC. This can be used to analyse the difference in error pattern between two acoustic parameters. Those regions where one feature performs well and the other one performs poorly should be further analysed to understand the underlying mechanism for the difference.

D. Error analysis

Some preliminary error analyses were performed at the EER thresholds for conditions 2 and 5. Similar analysis can be done in other operating points such as DCF.

1) The effect of age difference between two speakers on score of nontarget trial.

Fig. 13 shows the score box plots for each of the age differences between the two speakers in nontarget trials of condition 5. Similar plots can be obtained for condition 2 as well. It can be seen that the medians specified by the red dots decrease monotonically with the increase of age difference. This is more significant in female trials than in male trials. There are some outliers where the age difference is more than 50 years old, and this is probably due to the fewer number of trials in this range.

Fig. 14 a) and b) shows the histogram of nontarget trial numbers, and the histogram of false alarm with regard to the

speaker age difference. Fig. 14 c) is the error rate at every age difference and it can be seen that the error rate decreases monotonically with the age difference. Fig. 14 d) that shows that 80% of the errors happens for trials with age difference less than 10 years, whereas only 53% of trials are included in this range.

2) Speakers involved in errors of miss detection and false alarm.

Fig. 15 shows the number of miss detection errors for each speaker id in conditions 2 and 5. It can be seen that only a small portion of the speakers are involved in miss detection error, 21% for condition 2 and 14.5% for condition 5. Those speakers might be candidates of animal 'goat' in [6].

Fig. 16 shows the false alarm error rates for each speaker id in models and those speakers with error rate larger than 3% are indicated in red color. They might be candidates of animal 'lamb' in [6]. Fig. 17 shows the false alarm error rates for each speaker id in test segments and those speakers with error rates larger than 3% are candidates of animal 'wolf' [6]. However in PLDA system, the scores are symmetrical between model segment and test segment, so ideally the 'wolf' candidate group and the 'lamb' candidate group should be the same. This was proved by checking the overlapping between these two groups. For example, in condition 2, there are more than 100 speakers overlapped between these two groups.

3) Errors in various channels

Fig. 18, 19 and 20 show the number of errors along with the number of trials among channels for conditions 1, 2, and 5, respectively.

As shown in Fig. 18 for condition 1, mic 04 has many more errors than in mic 08, even though both have same number of trials. Further analysis is needed to understand this difference.

As shown in Fig. 19 for condition 2, mic 13 has more miss detection errors than other mics. Based on [13], this is a microphone array which is far from the speaker in the recording room. Correspondingly, mic 13 has a relatively lower false alarm than mic 5, 7, 8 and 12. However mic 02 has the lowest false alarm rate. Based on the error numbers in mic 07, mic 07 might be more similar to mic 08 than to mic 04.

As shown in Fig. 20 for condition 5, trials related to cell phone are dominant in both trial and error numbers. It can be seen that most errors happens when the two channels are not matched in terms of three phone types.

IV. SUMMARY

In this preliminary study, we took advantage of an information visualization and exploration tool (TIBCO Spotfire) to facilitate NIST SRE10 data visualization, exploration, and score and error analyses. All the key information provided by NIST for speakers, segments, and trials in the NIST SRE 2010 extended-core task was integrated together, along with the scores from our Probabilistic Linear Discriminant Analysis (PLDA) speaker

recognition system. We are able to conveniently visualize and exploit various types of information (such as speaker information, channel information, and scores for each trial) in many different ways (scatter plot, box plot, graph, etc.). In addition, the corresponding segments and voice activity detection (VAD) results for each trial are automatically linked to the trial table for inspection and comparison. Some mislabelled data was found through data visualization.

Our preliminary error analysis shows that the miss detection errors are from only a small portion (about 20% or less) of the speakers in trials, whereas most of the speakers in the database are involved in false alarms. The average score for the non-target trials decreases monotonically with the age difference between the two speakers, and most of the false alarms (about 80%) are involved in those trials where the age difference of two speakers is less than ten years old. We believe that further error analysis, facilitated by visualization and exploration tools, will provide us more insights in better understanding the speaker recognition systems and better designing the evaluation database. Those insights will help us bridge the gap between the dominant data-driven approaches in speaker recognition community and the scientific knowledge in speech science about speaker characteristics, and eventually improve the system performance.

A more thorough and more detailed analysis on the error patterns is in our plan of work.

REFERENCES

- [1] T. Kinnunen, and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, JAN 2010, 2010.
- [2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, Jan-Jul, 2000.
- [3] P. Kenny, N. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 16:980-988, 2008.
- [4] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV'07*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1-8.
- [5] Doddington, G. et al., 1998. Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: *Proc. ICSLP '98*.
- [6] G. Doddington, M. Przybocki, A. Martin *et al.*, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225-254, JUN 2000, 2000.
- [7] SPOTFIRE. TIBCO. <http://spotfire.tibco.com/>.
- [8] F. Nolan, *The phonetic bases of speaker recognition*, Cambridge [Cambridgeshire] ; New York: Cambridge University Press, 1983.
- [9] J. Kreiman, and D. Sidtis, *Foundations of voice studies: an interdisciplinary approach to voice production and perception*, Chichester, West Sussex ; Malden, MA: Wiley-Blackwell, 2011.
- [10] 2010 NIST Speaker Recognition Evaluation, http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf
- [11] D. Garcia-Romero and C. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems", *INTERSPEECH* 2011.
- [12] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, "Linear versus Mel-Frequency cepstral coefficients for speaker recognition", *ASRU 2011 (IEEE Automatic Speech Recognition and Understanding Workshop)*.
- [13] C. Christopher, L. Corson, D. Graff, K. Walker, "Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora", *Interspeech 2007*, Antwerp, August 2007.

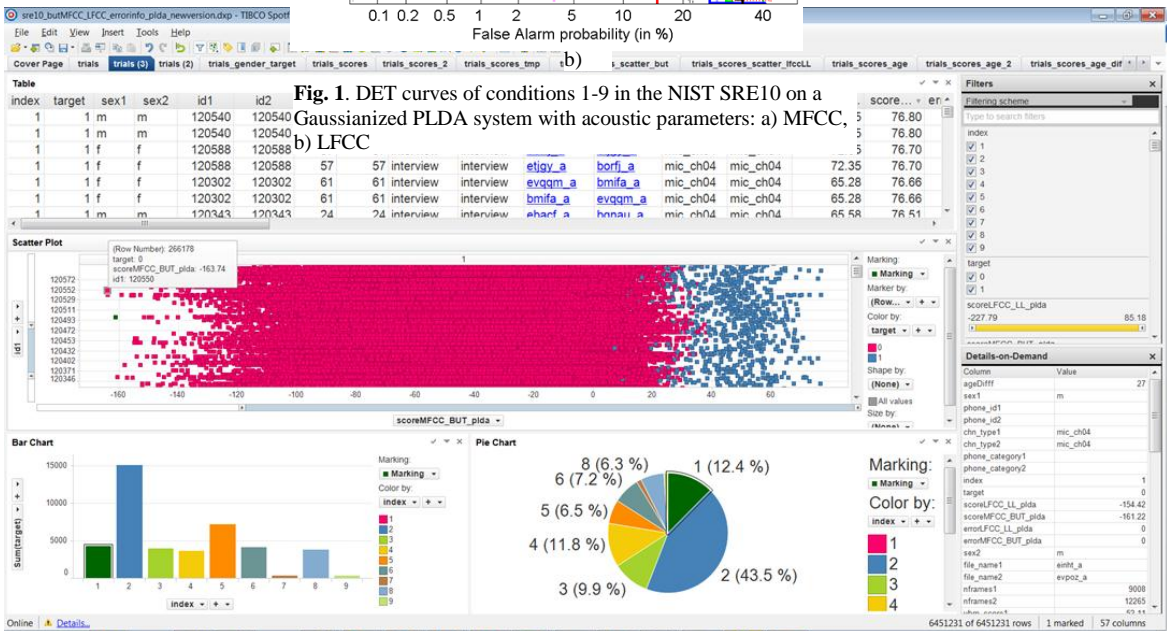
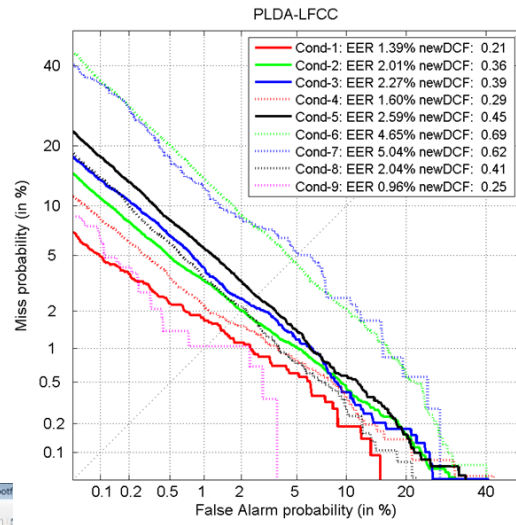
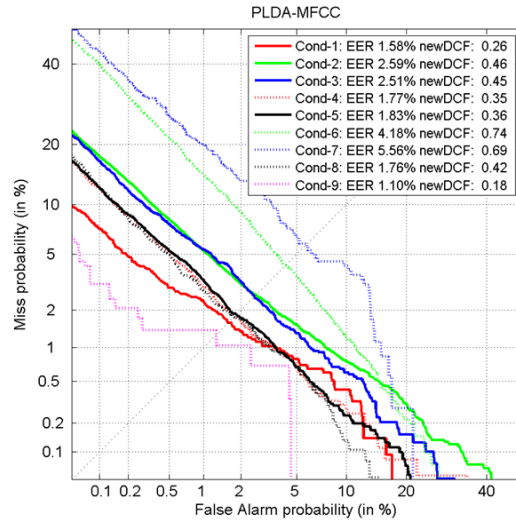


Fig. 2. A GUI screen snapshot of SPOTFIRE showing various windows for trial table, scatter plot, bar chart, pie chart, filter and details-on-demand)

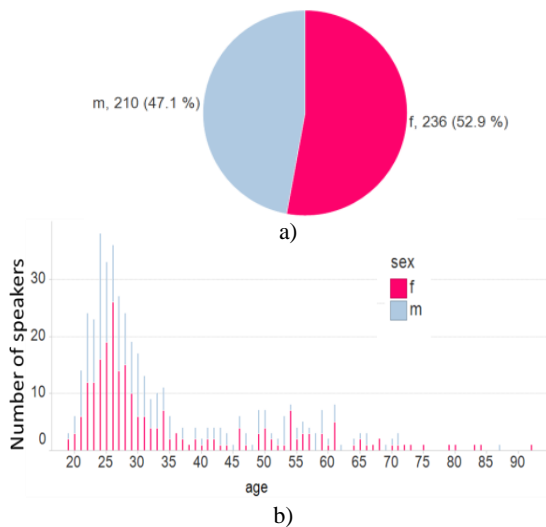


Fig. 3. Distribution of speakers (total 446) in the NIST SRE10. A) gender, B) age (modes: M 24, F 26). (m: male, f: female)

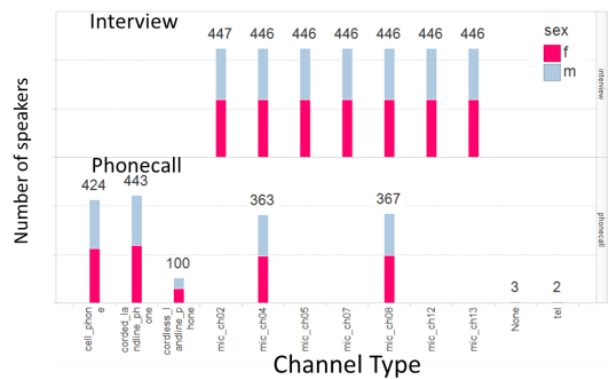
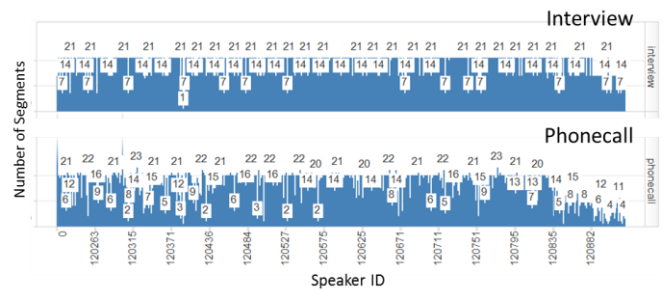
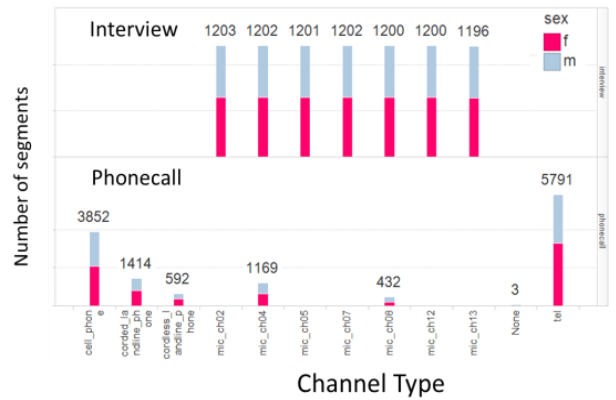
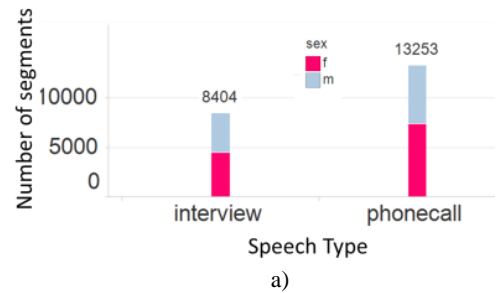


Fig. 4. Distribution of segments. A) Speech type, B) Channel type (from left to right: cell phone, corded landline, cordless landline, mic02, 04, 05, 07, 08, 12 and 13, "None" and "Tel" mean UNKNOWN channel), C) Number of segments per speaker (m: male, f: female), D) Number of speakers for each type of channel.

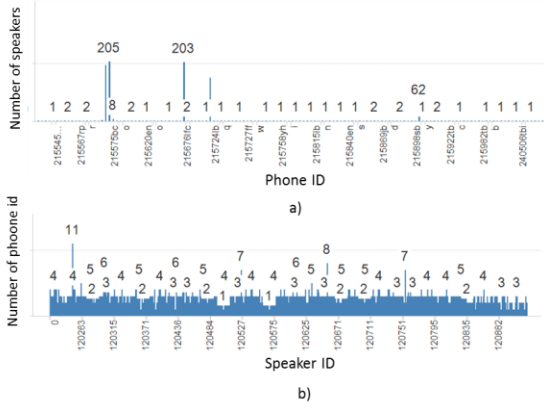


Fig. 5. Numbers of phone ID and speaker ID in the phonecall data. A) Number of speakers per phone ID, B) number of phone ID per speaker

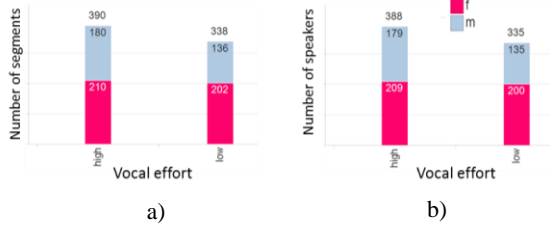


Fig. 6. Segments in high and low vocal effort. A) Number of segments, B) Number of speakers (m: male, f: female). Segments in high and low vocal effort only exist in corded land-line phone with only three corded-landline telephone IDs in total.

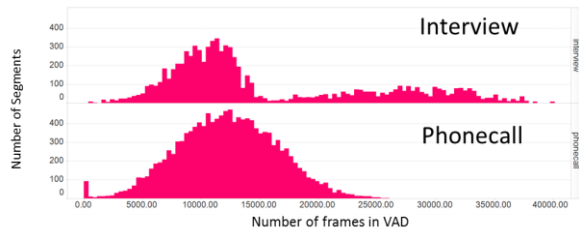


Fig. 7. Histograms of number of speech frames in voice activity detection (VAD)

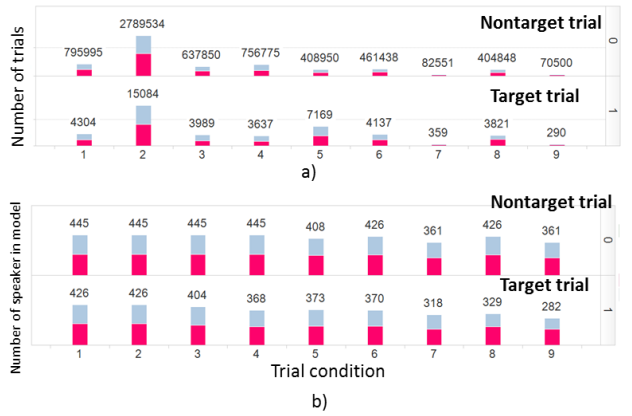


Fig. 8. Trials in condition 21-9 in the NIST SRE10 extended-core task. A) Number of trials in each condition, B). Number of speaker in model segments in each condition

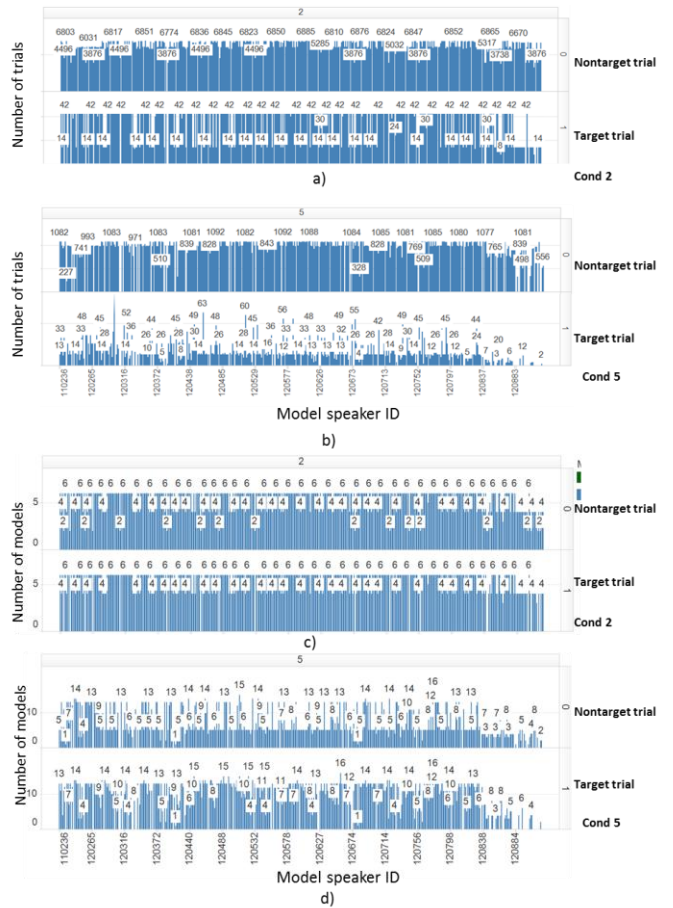


Fig. 9. Trial number and model number for each speaker id. A) Trial number in Condition 2, B) Trial number in Condition 5. C) Model number in Condition 2, D) Model number in Condition 5.

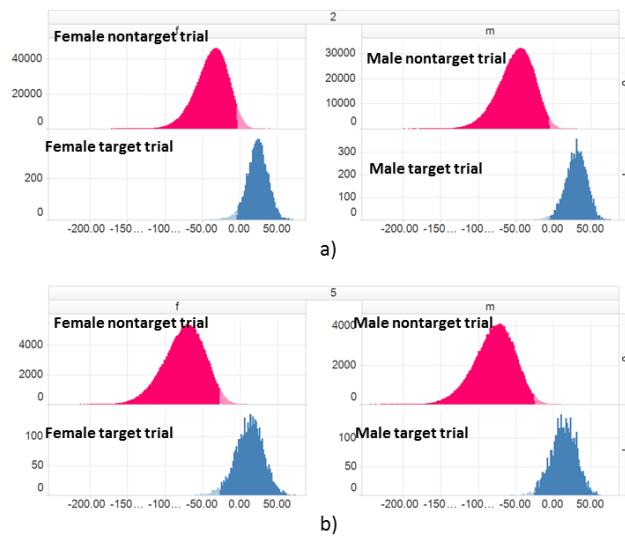


Fig. 10. Score histograms of trials (MFCC). A) Condition 2, B) Condition 5. In each plot, left: female, right: male, upper: nontarget trial, lower: target trial. (the light red/blue color indicates the trial errors for the EER threshold)

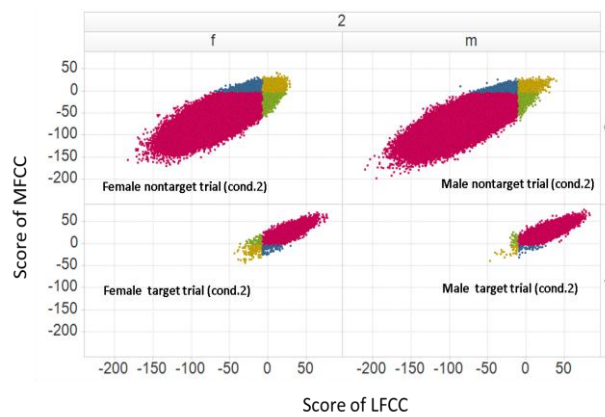


Fig. 12. Scatter plots of scores from both LFCC and MFCC (The color boundaries are for the threshold score values for EERs)

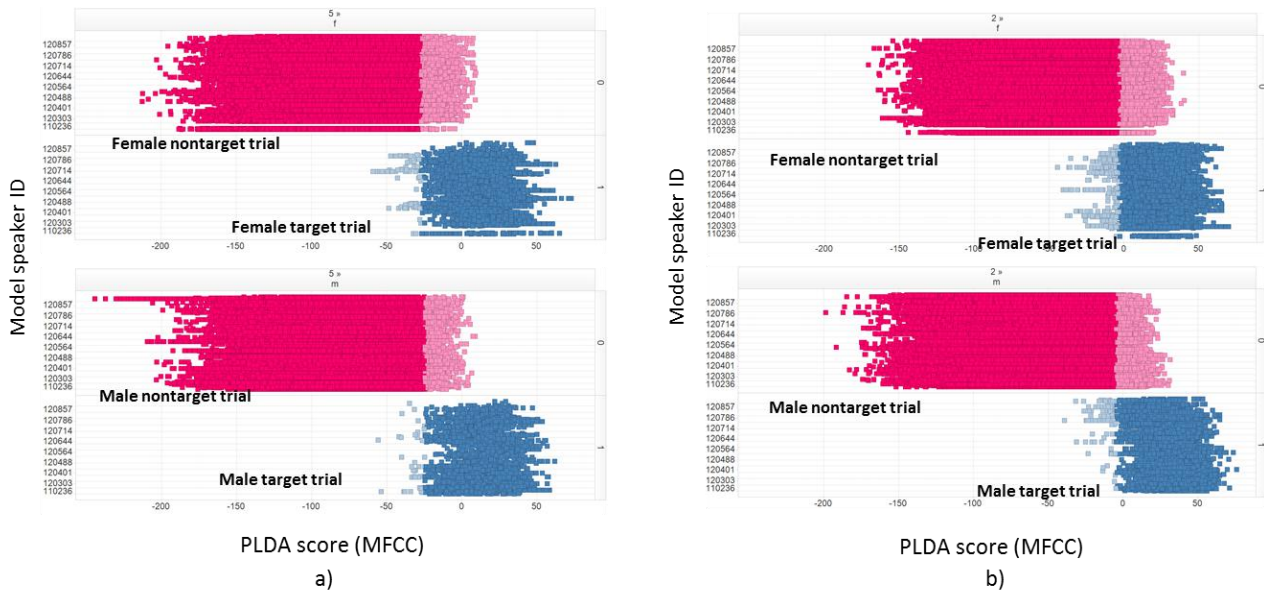


Fig. 11. Scatter plots of trial scores (MFCC). In each plot, left: Condition 5, right: Condition 2, upper: female, lower: male. Red color: nontarget trial, Blue color: target trial (the light red/blue color indicates the trial errors)

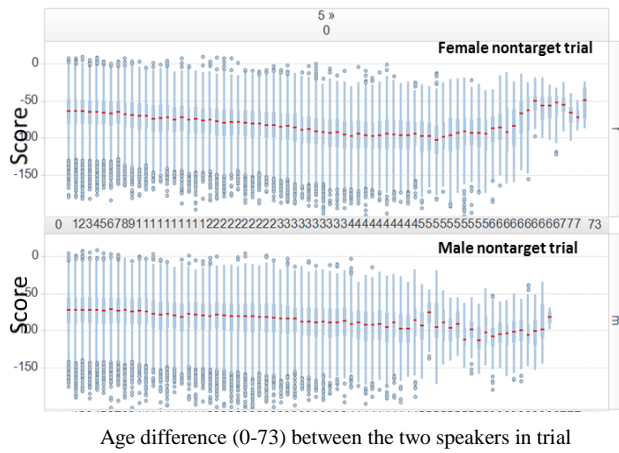


Fig. 13. Score box plots (condition 5) for each of the age difference between two speakers in each non-target trial (MFCC). (red dots are for medians) Upper panel: Female nontarget trial, lower panel: Male nontarget trial.

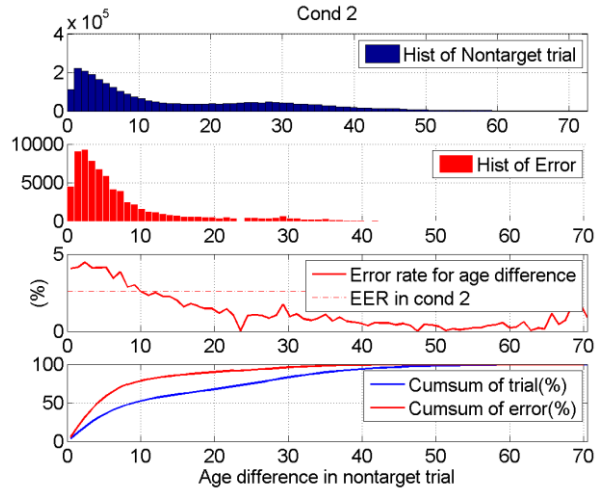


Fig. 14. Trial and error numbers in Condition 2 for the age difference between two speakers in non-target trials. A) Histogram of trial numbers. B) Histogram of error numbers (at EER threshold with MFCC). C) Error rate for each age difference. D) Cumulative trials number and error numbers in percentage.

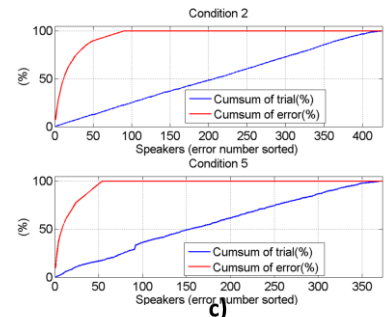


Fig. 15. Miss detection in conditions 2 and 5 at EER thresholds . a) Number of errors per speaker, b) Red: the number of speakers who have errors, Blue: all other speakers, c) the cumulative number of trials and errors.

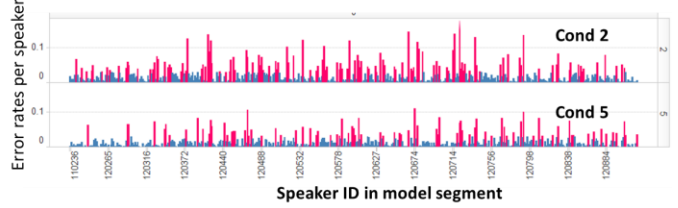


Fig. 16. False alarms in conditions 2 and 5 at EER thresholds a) Error rate per speaker ID in model segment, b) Red: the number of speakers with error rate $\geq 3\%$

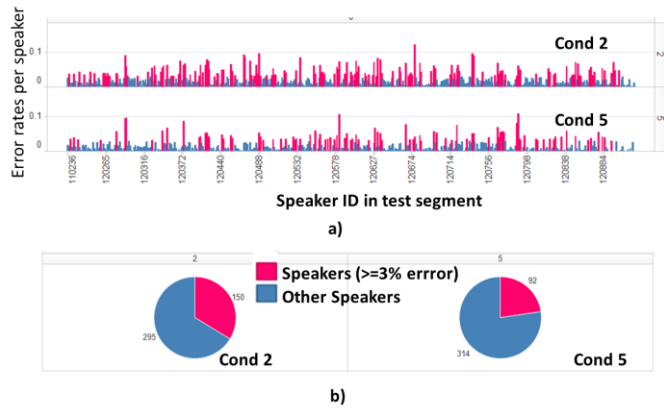


Fig. 17. False alarms in conditions 2 and 5 at EER thresholds a) Error rate per speaker ID in test segment, b) Red: the number of speakers with error rate $\geq 3\%$

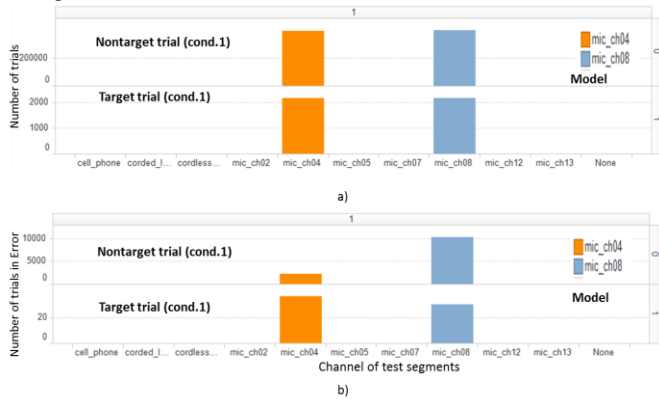


Fig. 18. Trial and error information in Condition 1 for different channels of test segments A) trial number . B) Error number



Fig. 19. Trial and error information in Condition 2 for different channels of test segments. A) trial number . B) Error number

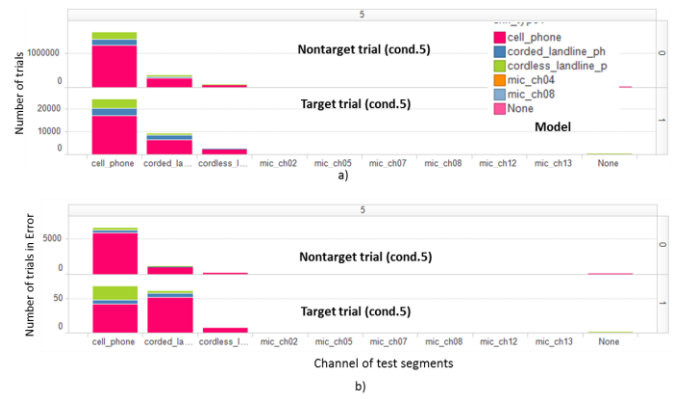


Fig. 20. Trial and error information in Condition 5 for different channels of test segments A) trial number . B) Error number