Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Hunting for Wolves in Speaker Recognition

Lara Stoll[1,2]*, George Doddington*

[1]International Computer Science Institute, Berkeley, CA, USA
[2]University of California at Berkeley, CA, USA
lstoll@icsi.berkeley.edu, george.doddington@gmail.com

Abstract

Identification and selection of speaker pairs that are difficult to distinguish offers the possibility of better focusing speaker recognition research, while also reducing the amount of data needed to estimate system performance with confidence. This work aims to predict which speaker pairs will be difficult for automatic speaker recognition systems to distinguish, by using features that characterize speakers, and thus provide a measure of speaker similarity. Features tested include pitch, jitter, shimmer, formant frequencies, energy, long term average spectrum energy, histograms of frequencies from roots of LPC coefficients, and spectral slope. Absolute and percent differences, Euclidean distance, and correlation coefficients are utilized to measure the closeness of these speaker features. Using data from NIST's 2008 Speaker Recognition Evaluation, the largest changes in detection cost and false alarm rate for similar speaker pairs (relative to all speaker pairs) occurs when speaker pairs are selected using the Euclidean distance between vectors of the mean first, second, and third formant frequencies. Even bigger differences in performance occur when speaker pairs are selected using the KL divergence between speaker-specific GMMs as a measure of similarity. In general, the feature-measures considered here are more successful at finding easy-to-distinguish speaker pairs than difficult-to-distinguish ones, and can provide potentially useful information about a speaker's tendency to be similar or dissimilar to other speakers.

1. Introduction

Automatic speaker recognition system performance depends on a variety of factors, including intrinsic speaker characteristics. As human listeners, we may observe that some speakers sound more alike than others. Similarly, we expect that automatic speaker recognition systems will vary across speaker pairs in how successfully each impostor speaker pair is classified correctly. By comparing the performance for a given speaker pair to performance over all speaker pairs, one can determine which speaker pairs are most (or least) difficult for a given system. Preliminary experiments indicated that poorly performing speaker pairs for one system generally performed poorly for other systems, suggesting that there are certain speaker pairs that are inherently difficult for any system. Rather than relying on a particular speaker recognition system's output to select such speaker pairs, we aim to find the universally difficult-to-distinguish speaker pairs by utilizing a variety of features, such as pitch, formant frequencies, or energy. Besides presenting an interesting problem, the results of such work may shed light on what speaker characteristics most affect speaker performance. If the speaker pairs most likely to cause errors can be identified, we may be able to use that knowledge and modify the system in order to improve overall performance. Another possible application of this work would be as a tool for NIST to select more difficult trials for future Speaker Recognition Evaluations, in order to present an even more challenging task.

There is some prior work demonstrating that automatic speaker recognition system performance depends on the speakers. Doddington et al. categorized speakers based on system performance [1]. Their study distinguished among speakers who cause a large number of false rejections as target speakers ("goats"), those who cause a large number of false acceptances as target speakers ("lambs"), those who cause a large number of false acceptances as impostor speakers ("wolves"), and default (well-behaved) speakers ("sheep"). In another article, Doddington et al. noted performance differences between high- and low-pitched speakers [2]. Poh et al. developed a user-specific score normalization (referred to as F-norm's variant) in order to address users who degrade system performance [3]. For a closed-set speaker identification task, Jin and Waibel implemented a method to reduce the effects of speakers who were likely to be identified as another speaker [4].

There is also a variety of related work in which prosodic and other features are used for speaker recognition or to otherwise characterize speaker differences. Speaker recognition approaches have used features like pitch and energy distributions or dynamics [5], prosodic statistics including duration and pitch-related features [6], and jitter and shimmer [7]. Formant frequencies and bandwidths, obtained using linear predictive (LP) analysis, were used as descriptors for perceptual speaker characterization by Necioğlu et al. [8], while McDougall and Nolan showed that formant frequency dynamics are speaker discriminative [9]. Kuwabara and Sagisaka considered many acoustic parameters as influences upon voice individuality, including pitch frequency, contour and fluctuation, formant frequencies, trajectories and bandwidths, and long-term average spectrum (LTAS) [10].

For our investigation, we consider a basic set of features, including pitch frequency statistics, energy statistics, LTAS energy statistics, formant frequency statistics, histograms of frequencies obtained from LP analysis, and spectral slope statistics. These features, along with appropriate distance measures, are utilized as a way to select speaker pairs that are closer, or more similar (in terms of that feature-measure pair). The goal is to find feature-measures for which similar speaker pairs correspond to speaker pairs that are difficult for automatic speaker recognition systems to distinguish. Furthermore, we also test the approximated Kullback-Liebler (KL) divergence between speaker-adapted Gaussian mixture models (trained on MFCC features) to provide a more complex measure that may better predict speaker recognition system behavior.

Section 2 describes the features, methods, and data used in our approach. The results of our experiments are given in Section 3. Finally, conclusions and future work are discussed in

Section 4.

2. Approach

Our approach tests a variety of measures calculated from different features as a criterion for selecting similar (or dissimilar) speaker pairs for speaker recognition. We describe the features considered in Section 2.1, and the measures and process of speaker pair selection are discussed in Section 2.2. The data used is covered in Section 2.3.

2.1. Features

The features described below are examined as potentially useful for speaker pair selection. Features are calculated either using MATLAB, and the Voicebox toolkit [11], or using Praat [12].

1. Pitch statistics (Praat): mean, median, range, and mean average slope of the pitch [f0_mean, f0_med, f0_range, f0_mas]
2. Jitter and shimmer (Praat): jitter relative average perturbation, and shimmer 5 point amplitude perturbation quotient [jitt_rap, shim_apq5]
3. Formant frequency statistics (Praat): mean and median of the first three formants [f1_mean, f1_med, f2_mean, f2_med, f3_mean, f3_med]
4. Energy statistics (Praat): mean and median energy [en_mean, en_med]
5. Long term average spectrum energy statistics (Praat): mean, standard deviation, range, slope, and local peak height of LTAS energy [ltas_mean, ltas_stddev, ltas_range, ltas_slope, ltas_lph]
6. Histograms of frequencies from roots of LPC coefficients (MATLAB/Voicebox): frequencies obtained from LPC order 14 coefficient roots (both with and without a minimum magnitude requirement of 0.88[1]) contribute to a histogram with a bin size of 5 Hz covering the 5-3995 Hz range [hist14all, hist14minmag]
7. Spectral slope statistics (MATLAB): mode and median of spectral slope, calculated over frequency range 0-4000 Hz [mode_specsl, med_specsl]

2.2. Measures and speaker pair selection

Features are calculated for each speech sample, and a measure is computed for every unique speaker pair in two ways. First is to average the feature values over all conversation sides of each speaker, and then calculate the measure for each speaker pair using these average per-speaker feature values. The second method calculates a measure for each possible pairing of conversation sides for a given speaker pair (with one conversation side for each speaker), and then averages these measure values to obtain a single value for each unique speaker pair. For each feature-measure, the results presented in Section 3 correspond to the most successful method.

For scalar features, absolute difference [absdiff] and percent difference [pctdiff] are used as measures, where percent difference for values x and y is defined as

$$\text{Percent difference} = \frac{|x-y|}{\frac{(x+y)}{2}}, \quad (1)$$

[1]This value was chosen based on a preliminary inspection of histograms, and was not optimized for selecting speaker pairs.

when x and y have the same sign (it is not used for features with both positive and negative values). In addition to the individual formants, sums of formants are used as scalar features (with absolute and percent difference measures), and the Euclidean distance [eucldist] is also calculated for vectors of formant frequencies, e.g. (f1,f2,f3). For the histograms of frequencies from LP analysis, a correlation coefficient [corr] is calculated as a measure of similarity. Table 1 summarizes the possible feature-measure combinations, grouped according to feature type.

Feature group	Features	Measures
Pitch statistics	f0_mean f0_med f0_range f0_mas	absdiff pctdiff
Jitter and shimmer	jitt_rap shim_apq5	absdiff pctdiff
Formant statistics	f1_mean f1_med f2_mean f2_med f3_mean f3_med	absdiff pctdiff
Sum of formant frequencies	f1+f2+f3_med f1+f2+f3_mean	absdiff pctdiff
Formant frequency vectors	(f1, f2)_mean (f1, f3)_mean (f2, f3)_mean (f1, f2)_med (f1, f3)_med (f2, f3)_med (f1, f2, f3)_mean (f1, f2, f3)_med	eucldist
Energy statistics	en_mean en_med	absdiff pctdiff
LTAS energy statistics	ltas_mean ltas_stddev ltas_range ltas_slope ltas_lph	absdiff pctdiff
LPC frequency histograms	hist14all hist14minmag	corr
Spectral slope statistics	mode_specsl med_specsl	absdiff pctdiff

Table 1: *Feature and measure combinations.*

Based on the measure for each unique speaker pair, those pairs with the highest and lowest 1% (or 5%) of values are selected to determine if the measure of speaker similarity corresponds to the degree of difficulty for a speaker recognition system. For absolute difference, percent difference, and Euclidean distance, smaller values should indicate more similar speakers, while for correlation coefficients, higher values indicate greater speaker similarity.

2.3. Speech corpora

The 2008 NIST Speaker Recognition Evaluation (SRE08) includes a condition (short2-short3) which uses roughly 2.5-3 minutes of speech for both training and testing [13]. This speech is taken either from one side of a conversation between two people over the telephone (possibly recorded on a microphone), or from part of an interview recorded on a microphone (some interviewer speech may be present). Additional interview data was released for a followup evaluation experiment designed to further explore the new interview style of data collection.

2.3.1. *Corpus for feature-measure calculation*

Speech data from the followup evaluation is used to calculate features for the speakers. In particular, speech recorded on microphone 2 (a lavalier microphone placed on the subject) is used since it has good sound quality. These speaker features are then used in conjunction with a similarity measure in order to predict difficult- and easy-to-distinguish speaker pairs. The majority of speakers have four conversation sides used for the measure calculation (a small minority have three or five conversation sides).

2.3.2. *Corpus for evaluation of selected speaker pairs*

The data used to evaluate speaker-pair selection is different in several respects from the data used to perform the selection. Specifically, the selection data were collected in an interview, while the evaluation data were collected in either an interview or a telephone conversation. Also, the selection data were collected using a lavalier microphone, whereas the evaluation data were collected using a variety of microphones, including a telephone handset. Furthermore, the selection data does not overlap with evaulation data.

Speaker recognition system submissions from the SRE08 short2-short3 condition are used to compute performance on trials for the selected 1% (or 5%) of most and least similar speaker pairs. Of the 34 sites who shared their system submissions for the short2-short3 condition, 33 of these are used in our results. The total number of trials for short2-short3 (after removing trials for speakers not found in the selection data) is 55013, with 1815 unique impostor speaker pairs. When keeping 1% (or 19) of the speaker pairs, there are around 4000 trials on average, while 5% (or 91) of the speaker pairs corresponds to an average of roughly 11000 trials. When filtering trials for selected speaker pairs, we removed target trials of speakers not included in any of the selected speaker pairs.

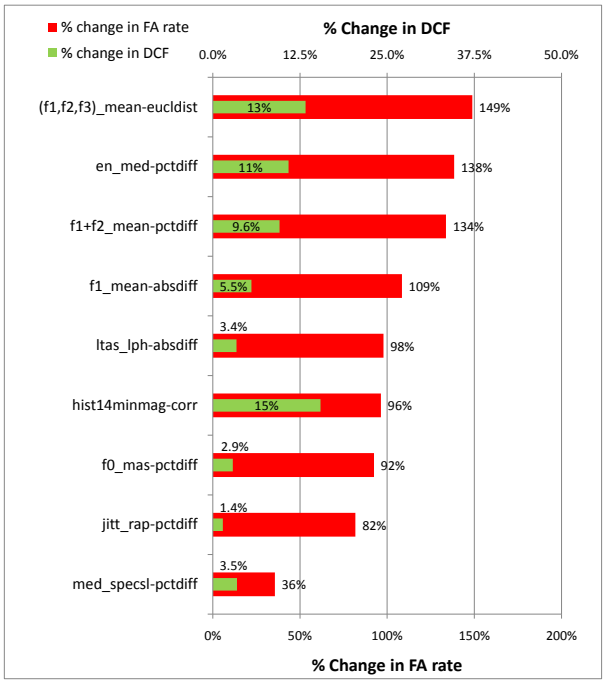

Figure 1: *Relative changes in DCF and FA rate for the most similar 1% of speaker pairs, compared to all speaker pairs.*

3. Results

System performance for the selected speaker pairs is reported using the minimum detection cost function (DCF) and false alarm (FA) rate, since we are concerned with finding difficult-to-distinguish impostor pairs. The DCF is defined as a weighted sum of the miss (i.e., not identifying a target speaker match) and false alarm (i.e., identifying an impostor speaker as the target speaker) error probabilities:

$$\begin{aligned}\mathrm{DCF} = {} & \mathrm{C}_{\mathrm{Miss}} \times \mathrm{P}_{\mathrm{Miss|Target}} \times \mathrm{P}_{\mathrm{Target}} + \\ & \mathrm{C}_{\mathrm{FalseAlarm}} \times \mathrm{P}_{\mathrm{FalseAlarm|NonTarget}} \times (1 - \mathrm{P}_{\mathrm{Target}}) \end{aligned} \quad (2)$$

In Equation (2), $\mathrm{C}_{\mathrm{Miss}}$ and $\mathrm{C}_{\mathrm{FalseAlarm}}$ are the relative costs of detection errors, and $\mathrm{P}_{\mathrm{Target}}$ is the *a priori* probability of the specified target speaker. SRE08 used $\mathrm{C}_{\mathrm{Miss}} = 10$, $\mathrm{C}_{\mathrm{FalseAlarm}} = 1$, and $\mathrm{P}_{\mathrm{Target}} = 0.01$. For a given decision threshold, the FA rate is defined as:

$$\mathrm{P}_{\mathrm{FalseAlarm}} = \frac{\text{number of false alarm errors}}{\text{total number of nontarget trials}} \quad (3)$$

For each speaker recognition system, we compute the change in minimum DCF for the most (and least) similar speaker pairs relative to all speaker pairs. Compared to a FA rate of 1% on all speaker pairs, we calculate the change in FA rate (at the decision threshold yielding 1% FA on all trials) for the most (and least) similar pairs. These relative differences are then averaged over all systems. With each feature-measure, if more similar (i.e., closer) speaker pairs correspond to difficult-to-distinguish speaker pairs, then changes in the DCF and FA rate should be positive and significant. The converse holds for less similar speaker pairs, which will have significant negative changes if they are easier for systems to distinguish.

Figure 2: *Relative changes in DCF and FA rate for the least similar 1% of speaker pairs, compared to all speaker pairs.*

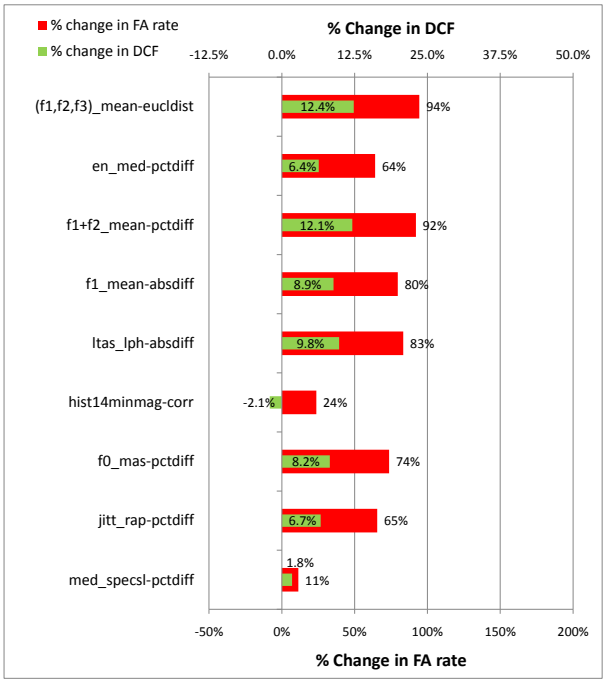

Figure 3: *Relative changes in DCF and FA rate for the most similar 5% of speaker pairs, compared to all speaker pairs.*

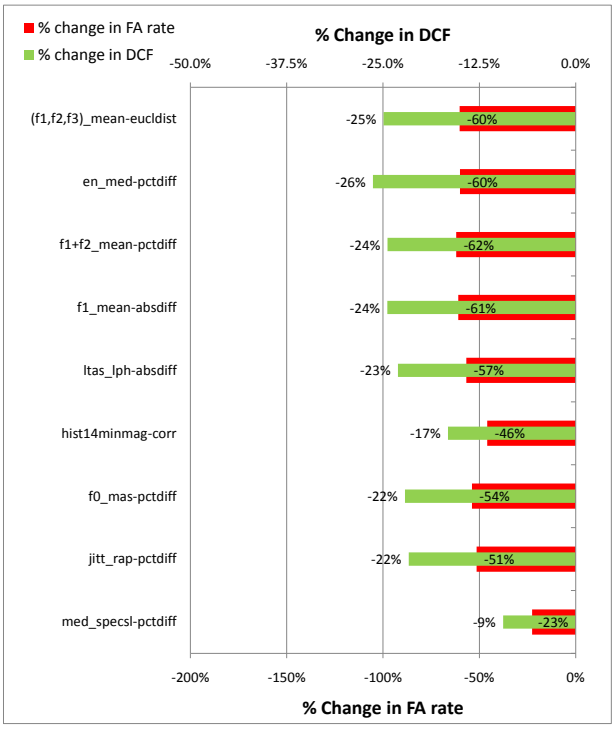

Figure 4: *Relative changes in DCF and FA rate for the least similar 5% of speaker pairs, compared to all speaker pairs.*

Figures 1 and 2 show performance differences for the top 1% most and least similar speaker pairs, respectively. For each feature group, the feature-measure pair yielding the largest DCF and FA changes is presented. Similarly, Figures 3 and 4 show results when considering the top 5% most and least similar speaker pairs, respectively.

We observe that features of each type can select speaker pairs for which the most (or least) similar have worse (or better) performance than all speaker pairs. Furthermore, this difference in performance typically increases when a smaller fraction of speaker pairs is used, i.e., there is a bigger difference for the most similar 1% of speaker pairs than for the most similar 5%. It should be noted that changes in performance are not uniform across different speaker verification systems.

The feature-measure that yields the largest average difference in performance for the 1% most similar speaker pairs is the Euclidean distance between vectors of the mean first, second, and third formant frequencies. The next best feature-measures include other formant-based measures, the percent difference of median energy, and the correlation of histograms of LPC freqencies with minimum magnitude requirement. For the 1% least similar speaker pairs, results are fairly similar across feature-measures, with the correlation of LPC frequency histograms and spectral slope yielding the smallest differences. The Euclidean distance between vectors of the mean, first, second, and third formant frequencies also appears to be the best feature-measures for finding the 5% most difficult-to-distinguish speaker pairs, with the percent difference of the sum of formants and the absolute difference in LTAS local peak height being the next best. As with the 1% least similar speaker pairs, the 5% least similar show very consistent results across feature-measures, with reduced effectiveness for the correlation of LPC frequency histograms and spectral slope.

Detection error tradeoff (DET) curves are shown for example systems in Figures 5 and 6, using the Euclidean distance between vectors of the means of the first, second and third formants, and the percent difference of the median energy, respectively. Although the system in Figure 6 has good separation among the different DET curves, we observe more overlap in the DET curves of Figure 5. Furthermore, Figure 5 reveals an asymmetry in behavior for dissimilar and similar speaker pairs, showing that the performance on difficult-to-distinguish speaker pairs is closer to performance on all speaker pairs. While this asymmetry does not exist for all systems and all sets of selected speaker pairs (as evidenced by Figure 6), the trend does hold in most cases.

While the results presented thus far do show some promise, the changes in performance for similar speaker pairs (relative to all speaker pairs) are not particularly large. Accordingly, we tested a measure that utilizes Gaussian mixture models, with the idea that GMMs may better predict speaker recognition system performance, given that many systems utilize cepstral feature-trained GMMs. Using SRI's tools for training GMMs for speaker recognition [14], we trained speaker-specific GMMs via maximum *a posteriori* (MAP) adaptation from a universal background model trained on Fisher data. The input features were 12th order MFCCs plus energy, with deltas and double-deltas, and the models used 1024 Gaussians. For each unique pair of speaker-specific GMMs, an approximation to the Kullback-Leibler (KL) divergence (based on the unscented transform [15]) was used to measure similarity. Results are shown in Figure 7.

Compared to previous feature-measures, the KL divergence is indeed more effective at finding difficult- and easy-to-distinguish speaker pairs. DET curves for an example system are shown in Figure 8. Again, we observe that, relative to per-

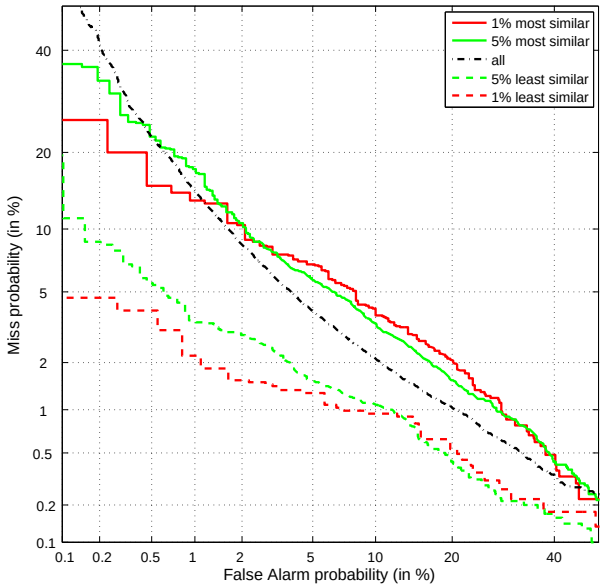

Figure 5: *DET curves for an illustrative speaker recognition system, using the Euclidean distance between vectors of the mean first, second, and third formant frequencies for speaker pair selection.*

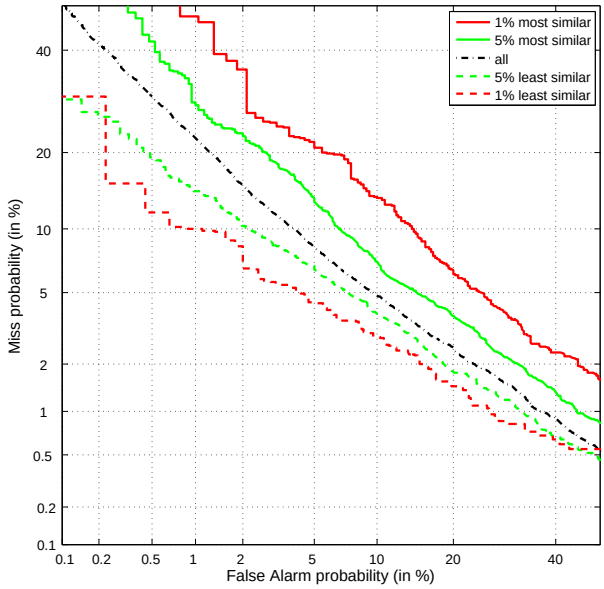

Figure 6: *DET curves for an illustrative speaker recognition system, using the percent difference of median energy for speaker pair selection.*

formance on all speaker pairs, there is a larger performance gap for dissimilar speaker pairs than for similar speaker pairs.

Our feature-measures seem to be more successful at selecting easy-to-distinguish speaker pairs, suggesting that these speaker pairs may be easier to find. Such a result is not surprising, considering that speaker pairs that are very dissimilar in terms of pitch, formant frequencies, or other such feature, are most likely different enough to not be confused by a speaker recognition system. On the other hand, these single features may be unable to capture the complexities of what makes a speaker pair hard for the system to distinguish. In other words, dissimilar speaker pairs are likely to differ largely in a number of characteristics, so that any one of the characteristics may be sufficient to identify them, while similarity in a single characteristic does not necessarily mean a speaker pair will be difficult-to-distinguish.

Returning to the groups of speaker pairs selected by the KL divergence approximation for GMMs, let us more closely examine the 1%, 3%, 5%, 10%, and 20% most and least similar speaker pairs. Overall, there are 150 speakers, with 87 female and 63 male, for which there are 1815 same-sex impostor speaker pairs with impostor trials in the SRE08 short2-short3 task. For the groups of speaker pairs with larger values for KL divergence, that is, those speaker pairs that are expected to be easier for systems to distinguish, the majority are male (close to 75% on average). The opposite tendency holds to a lesser extent for more similar pairs tending to be female, although the groups with the lowest 1% and 3% of KL divergence values still have more male speaker pairs. These results suggest that there is a greater range of differences among male speakers, so that there are likely to be more dissimilar male speaker pairs.

Furthermore, examining the number of times a particular speaker appears in a group of similar or dissimilar speaker pairs, we note that there tend to be two types of speakers: those who appear frequently as members of difficult-to-distinguish speaker pairs, and those who occur frequently as members of easy-to-distinguish speaker pairs. In fact, there are 15 speakers (1 male, 14 female) that never appear in the most-dissimilar groups, and 24 speakers (10 male, 14 female) that never appear in the most-similar groups. Such a result supports the existence of the "wolves" and "lambs" described in the aforementioned work of Doddington et al. [1].

4. Conclusions and future work

In summary, the results of this work demonstrate that, to a certain extent, it is possible to predict which speaker pairs will be difficult for a typical speaker recognition system to distinguish. Both difficult- and easy-to-distinguish speaker pairs can be selected using a measure of similarity calculated from features like pitch, energy, or spectral slope. For the features considered here, using the Euclidean distance between vectors of mean first, second, and third formant frequencies produces the largest difference in performance for similar and dissimilar speaker pairs. An even more successful measure is the KL divergence calculated between speaker-specific GMMs. Overall, the degree of success is higher for selecting dissimilar speaker pairs than it is for selecting similar speaker pairs, possibly because similarity in a single characteristic is not necessarily sufficient to identify a difficult-to-distinguish speaker pair. While the feature-measures may not be as effective at finding difficult-to-distinguish speaker pairs as desired, they still provide potentially useful information about speakers. In particular, one may be able to determine an overall tendency of a speaker to be similar or dissimilar to other speakers.

In terms of future work, one task is to test combinations of multiple feature-measure pairs, since it may be possible to improve success in finding similar speaker pairs if multiple feature-measures are used as selection criteria. Another task is to extend this work to find features for selecting target speakers that are difficult for the system to correctly recognize (in other

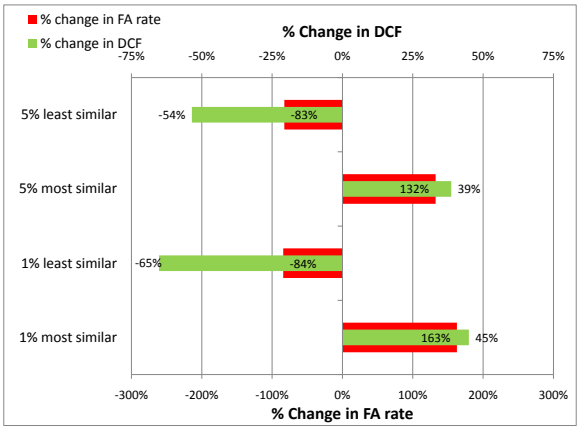

Figure 7: *Relative changes in DCF and FA rate for the most and least similar 1% and 5% of speaker pairs selected by the approximated KL divergence between speaker-specific GMMs.*

words, to consider the behavior of speakers for target rather than impostor trials). Given the lack of consistency in how different systems behave for the same set of speakers, further investigations may be able to reveal trends in behavior across classes or types of systems.

5. Acknowledgements

This material is based upon work supported by NSF under grant number 0329258. Thank you to Eduardo Lopez, who was kind enough to share some of his code. Finally, thanks to NIST and all the sites who shared their SRE08 system submissions.

6. References

[1] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. ICSLP*, 1998.

[2] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225 – 254, 2000.

[3] Norman Poh, Samy Bengio, and Arun Ross, "Revisiting Doddington's zoo: A systematic method to assess user-dependent variabilities," in *Proc. of Multimodal User Authentication*, 2006.

[4] Qin Jin and Alex Waibel, "A naive de-lambing method for speaker identification," in *Proc. ICSLP*, 2000.

[5] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP*, 2003.

[6] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," in *Proc. ICASSP*, 2003.

[7] Mireia Farrús, Javier Hernando, and Pascual Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Proc. Interspeech*, 2007.

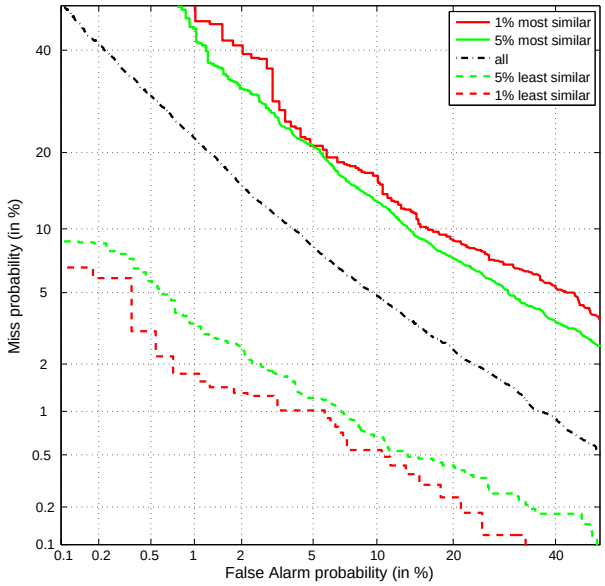

Figure 8: *DET curves for an illustrative speaker recognition system, using the approximated KL divergence between speaker-specific GMMs to select speaker pairs.*

[8] Burhan F. Necioğlu, Mark A. Clements, and Thomas P. Barnwell III, "Objectively measured descriptors applied to speaker characterization," in *Proc. ICASSP*, 1996.

[9] Kirsty McDougall and Francis Nolan, "Discrimination of speakers using the formant dynamics of /u:/ in british english," in *Proc. ICPhS*, J. Trouvain and W. Barry, Eds., 2007, pp. 1825–1828.

[10] Hisao Kuwabara and Yoshinori Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, pp. 165–173, 1995.

[11] Mike Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[12] Paul Boersma and David Weenink, "Praat: doing phonetics by computer (version 5.0.3.0)," http://www.praat.org.

[13] National Institute of Standards and Technology, "The NIST year 2008 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf, 2008.

[14] Sachin S. Kajarekar, Luciana Ferrer, Elizabeth Shriberg, Kemal Sonmez, Andreas Stolcke, Anand Venkataraman, and Jing Zheng, "SRI's 2004 NIST speaker recognition evaluation system," in *Proc. ICASSP*, 2005.

[15] Jacob Goldberger and Hagai Aronowitz, "A distance measure between gmms based on the unscented transform and its application to speaker recognition," in *Proc. Eurospeech*, 2005.