

# Hunting for Wolves in Speaker Recognition

Lara Stoll\*  
George Doddington



\*International Computer Science Institute  
\*University of California at Berkeley  
Dept. of Electrical Engineering and  
Computer Science



# Overview

---

- ▶ Objective
- ▶ Related work
- ▶ Approach
- ▶ Results
- ▶ Analysis
- ▶ Conclusions and future work




# Objective

---

- ▶ Automatic speaker recognition performance depends on a variety of factors, including intrinsic speaker characteristics
- ▶ Aim: To predict which (impostor) speaker pairs will be difficult for automatic speaker recognition systems to distinguish
  - ▶ Use features that provide a measure of speaker similarity (pitch statistics, formant frequency statistics, spectral slope, etc.)
- ▶ Motivation: better focus speaker recognition research and reduce amount of data needed to estimate system performance with confidence

# Related Work – Speaker types [1/2]

---

- ▶ Doddington et al. (1998) – categorized speakers based on system performance
  - ▶ Goats – cause a large number of false rejections as target speakers 
  - ▶ Lambs – cause a large number of false acceptances as target speakers  
  - ▶ Wolves – cause a large number of false acceptances as impostor speakers
  - ▶ Sheep – default, well-behaved speakers 

## Related Work – Speaker types [2/2]

---

- ▶ Doddington et al. (2000) – found performance differences between high- and low-pitched speakers
- ▶ Poh et al. (2006) – user-specific score normalization to address users degrading system performance
- ▶ Jin and Waibel (2000) – method for reducing effects of speakers who were likely to be identified as another speaker applied to closed-set speaker identification task

# Related Work – Speaker features

---

- ▶ **Speaker recognition approaches**
  - ▶ Pitch and energy distributions or dynamics
  - ▶ Prosodic statistics including duration and pitch-related features
  - ▶ Jitter and shimmer
- ▶ **Perceptual speaker characterization/discrimination**
  - ▶ Formant frequencies and bandwidths
  - ▶ Formant frequency dynamics
- ▶ **Acoustic parameters influencing voice individuality**
  - ▶ Pitch frequency, contour and fluctuation
  - ▶ Formant frequencies, trajectories and bandwidths
  - ▶ Long-term average spectrum (LTAS)

# Approach - Overview

---

- ▶ Compute a feature for each conversation side
- ▶ Compute measure of similarity (using feature values) for each impostor speaker pair
- ▶ Select speaker pairs with the highest and lowest 1% (or 5%) of values and compare performance on these speaker pairs to performance on all speaker pairs

# Approach - Features

---

- ▶ Pitch frequency statistics: mean, median, range, and mean average slope



# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitter relative average perturbation and shimmer 5 point amplitude perturbation quotient

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: mean and median of f1, f2, and f3

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: f1\_mean, f1\_med, f2\_mean, f2\_med, f3\_mean, f3\_med
- ▶ Energy statistics: mean and median energy

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: f1\_mean, f1\_med, f2\_mean, f2\_med, f3\_mean, f3\_med
- ▶ Energy statistics: en\_mean, en\_med
- ▶ LTAS energy statistics: mean, standard deviation, range, slope, and local peak height of LTAS energy

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: f1\_mean, f1\_med, f2\_mean, f2\_med, f3\_mean, f3\_med
- ▶ Energy statistics: en\_mean, en\_med
- ▶ LTAS energy statistics: ltas\_mean, ltas\_stddev, ltas\_range, ltas\_slope, ltas\_lph

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: f1\_mean, f1\_med, f2\_mean, f2\_med, f3\_mean, f3\_med
- ▶ Energy statistics: en\_mean, en\_med
- ▶ LTAS energy statistics: ltas\_mean, ltas\_stddev, ltas\_range, ltas\_slope, ltas\_lph
- ▶ Histograms of frequencies from LPC: frequencies obtained from LPC order 14 coefficient roots, both with and without a minimum magnitude requirement of 0.88, contribute to a histogram with bin size of 5 Hz covering the 5-3995 Hz range

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: f1\_mean, f1\_med, f2\_mean, f2\_med, f3\_mean, f3\_med
- ▶ Energy statistics: en\_mean, en\_med
- ▶ LTAS energy statistics: ltas\_mean, ltas\_stddev, ltas\_range, ltas\_slope, ltas\_lph
- ▶ Histograms of frequencies from roots of LPC coefficients: hist14all, hist14minmag
- ▶ Spectral slope statistics: mode and median of spectral slope, calculated over frequency range 0-4000 Hz

# Approach - Features

---

- ▶ Pitch frequency statistics: f0\_mean, f0\_median, f0\_range, f0\_mas
- ▶ Jitter and shimmer: jitt\_rap, shim\_apq5
- ▶ Formant frequency statistics: f1\_mean, f1\_med, f2\_mean, f2\_med, f3\_mean, f3\_med
- ▶ Energy statistics: en\_mean, en\_med
- ▶ LTAS energy statistics: ltas\_mean, ltas\_stddev, ltas\_range, ltas\_slope, ltas\_lph
- ▶ Histograms of frequencies from roots of LPC coefficients: hist14all, hist14minmag
- ▶ Spectral slope statistics: mode\_specsl, med\_specsl



# Approach – Similarity measures

---

- ▶ Distance/similarity measures
  - ▶ For scalar features: absolute or percent difference
  - ▶ Vectors of formant frequency statistics: Euclidean distance
  - ▶ Histograms of frequencies: correlation
- ▶ Two ways of computing measure of speaker pair similarity:
  - ▶ Average feature over all conversation sides of a speaker, and then compute distance measure between average feature values for two speakers
  - ▶ Compute distance measure between features of each conversation pair of two speakers, and then average over these measures

# Approach – Corpora

---

- ▶ **Feature-measure calculation**
  - ▶ NIST SRE08 followup evaluation data
    - ▶ Interview data (lavalier microphone chosen for quality)
    - ▶ Majority of speakers have 4 conversation sides
- ▶ **Evaluation of selected speaker pairs**
  - ▶ NIST SRE08 short2-short3 data
    - ▶ Roughly 2.5-3 minutes of speech from telephone conversation (possibly recorded on a microphone) or interview (recorded on a microphone)

# NIST SRE2008 short2-short3 condition

---

- ▶ 33 primary submissions shared by participating sites
- ▶ After removing trials corresponding to speakers not found in the selection data, there are 55013 trials, with 1815 unique speaker pairs
- ▶ Keeping 1% (or 19) of speaker pairs leaves ~4000 trials on average
- ▶ Keeping 5% (or 91) of speaker pairs leaves ~11000 trials on average
- ▶ Target trials of speakers not included in any of the selected speaker pairs are removed

# Evaluation of system performance

---

- ▶ Minimum detection cost function (DCF):

- ▶  $DCF = C_{\text{Miss}} \times P_{\text{Miss}|\text{Target}} \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm}|\text{NonTarget}} \times (1 - P_{\text{Target}})$

- ▶ Relative costs of errors:  $C_{\text{Miss}} = 10, C_{\text{FalseAlarm}} = 1$

- ▶ *A priori* probability of target speaker:  $P_{\text{Target}} = 0.01$

- ▶ At a given decision threshold, the false alarm (FA) rate is:

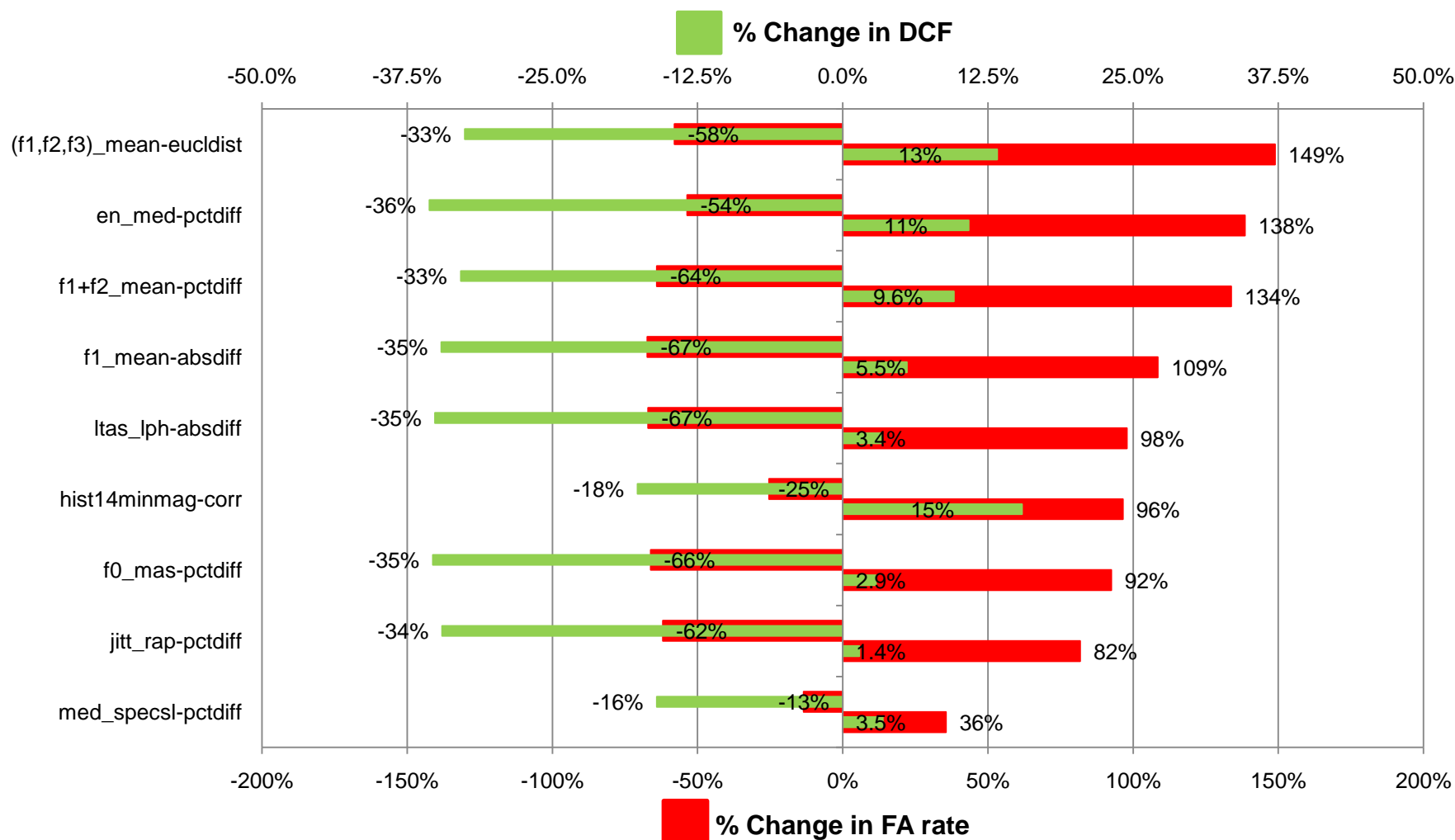
- ▶  $P_{\text{FalseAlarm}} = \frac{\text{Number of false alarm errors}}{\text{Total number of nontarget trials}}$

# Presentation of results

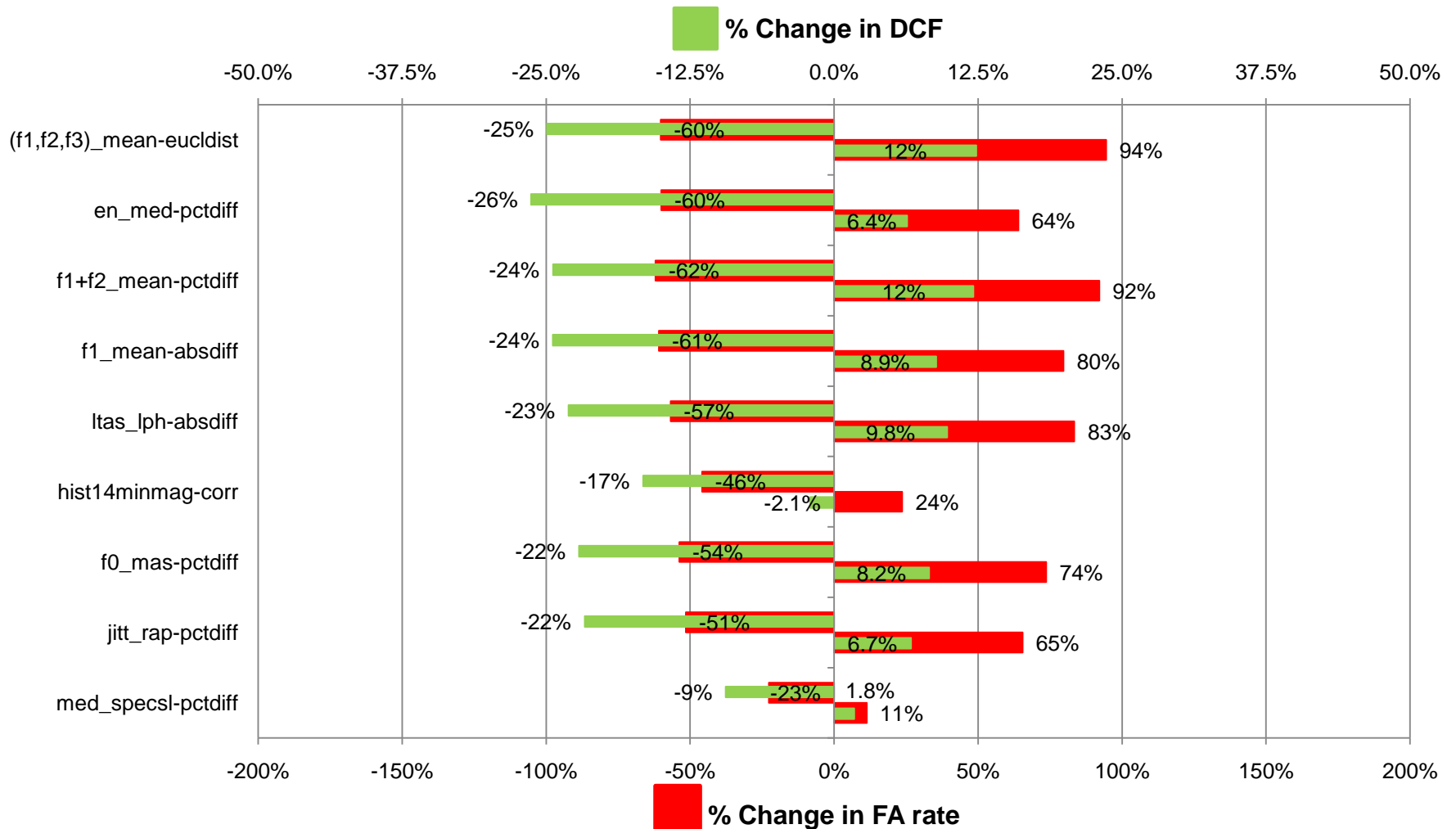
---

- ▶ For each system submission, compute the change in min DCF for the most (and least) similar speaker pairs relative to all speaker pairs; then average these changes over all systems
- ▶ Compared to FA rate of 1%, calculate change in FA rate (at same decision threshold that yields 1% FA on all trials) for most (and least) similar speaker pairs
- ▶ If more similar (according to a given feature-measure) corresponds to more difficult-to-distinguish, the changes in DCF and FA rate should be positive

# Results for 1% of speaker pairs



# Results for 5% of speaker pairs



# Observations/Comments

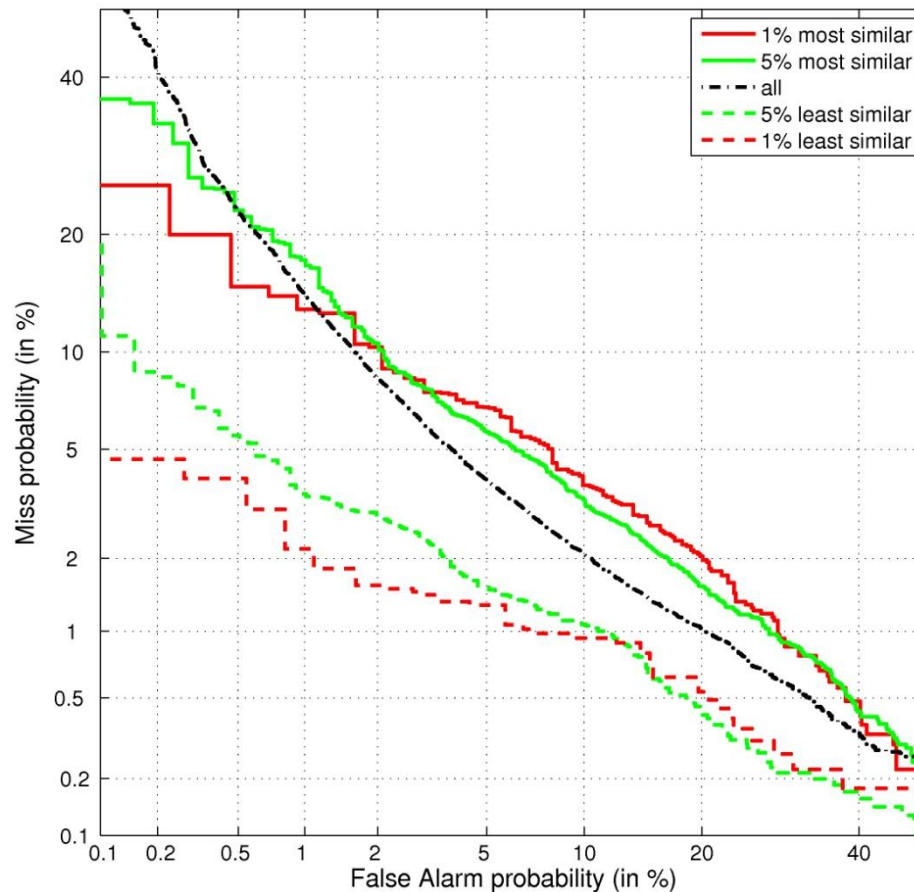
---

- ▶ Can successfully select speaker pairs for which most (or least) similar have worse (or better) performance than all speaker pairs
- ▶ Larger change in performance for top 1% most and least similar speaker pairs than top 5%
- ▶ Best feature-measure is the Euclidean distance between vectors of (f1\_mean, f2\_mean, f3\_mean)
- ▶ Note: changes in performance are not uniform across site submissions



# DET curves for illustrative system [1/2]

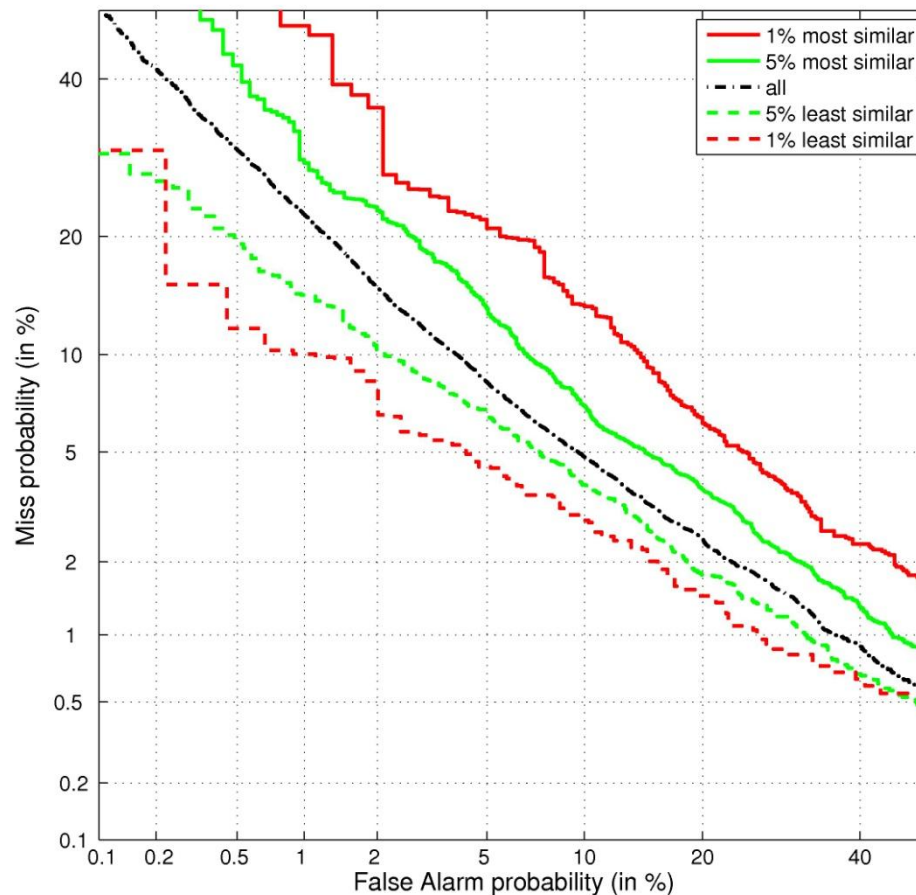
Feature-measure:  
Euclidean distance  
between  
vectors of  
the mean  
first, second,  
and third  
formant  
frequencies



Note:  
Asymmetry of  
behavior for  
dissimilar and  
similar  
speaker pairs;  
difficult-to-dist  
inguish curves  
are closer to  
all speakers  
curve – trend  
holds in many  
cases

# DET curves for illustrative system [2/2]

Feature-measure:  
Percent difference of median energy



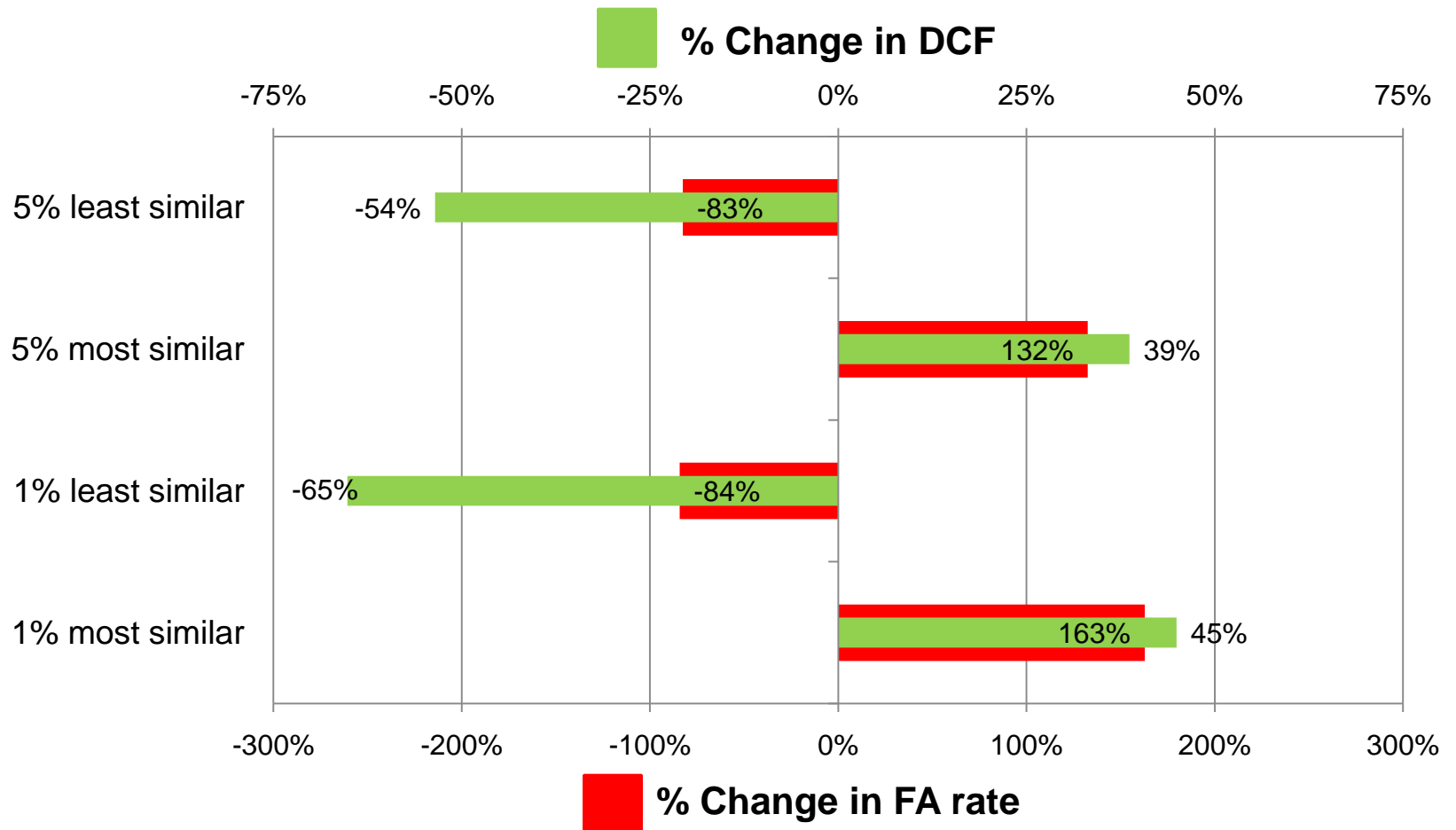
Note: Good separation of DET curves, unlike overlap shown previously

# Better measure for speaker pair selection

---

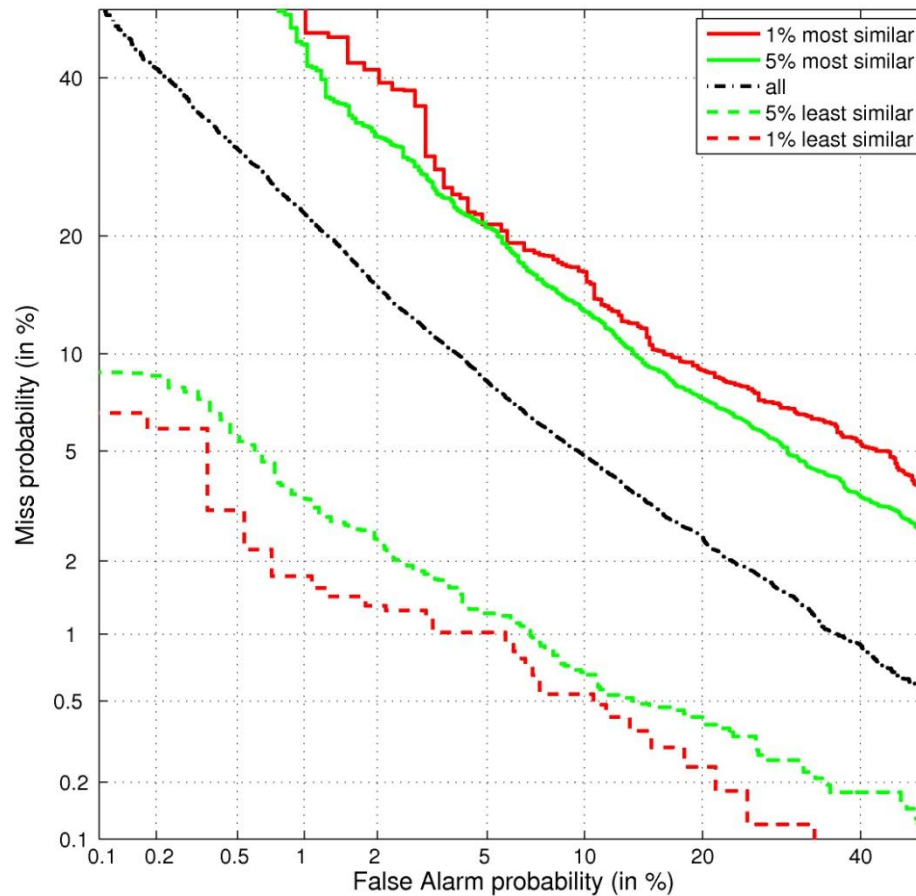
- ▶ Utilize GMMs (since many systems use cepstral feature-trained GMMs)
- ▶ Speaker-specific GMMs MAP adapted from a UBM (trained on Fisher data)
- ▶ 12<sup>th</sup> order MFCCs plus energy, with deltas and double-deltas
- ▶ 1024 Gaussians
- ▶ Similarity measure between speaker GMMs: an approximation to the Kullback-Leibler (KL) divergence based on the unscented transform

# Results: KL divergence between GMMs



# DET curves for illustrative system

Approximate  
KL divergence  
between  
speaker-  
specific GMMs



Note: Again we have a larger performance gap for dissimilar speaker pairs than similar speaker pairs, relative to all speaker pairs

## Additional Analysis/Breakdown of Results [1/3]

---

- ▶ Typically more successful at selecting easy-to-distinguish speaker pairs – these pairs may be easier to find
- ▶ Possible explanation:
  - ▶ Speaker pairs that are very dissimilar in terms of pitch, formant frequencies, etc., are most likely different enough to not be confused by the system
  - ▶ But, single features may be unable to capture the complexities of what makes a speaker pair hard to distinguish

## Additional Analysis/Breakdown of Results [2/3]

---

- ▶ For KL divergence measure, examine the 1%, 3%, 5%, 10%, and 20% most and least similar speaker pairs
- ▶ 150 speakers overall, 87 female and 63 male; 1815 same-sex impostor speaker pairs (with trials in SRE08 short2-short3 task)
  - ▶ Groups of speaker pairs with larger values of KL divergence (i.e., expected to be easy-to-distinguish): majority are male (~75% on average)
  - ▶ Opposite tendency holds to a lesser extent for more similar pairs tending to be female (lowest 1% and 3% of KL divergence values still have more male pairs)
  - ▶ Suggests that there is a greater range of differences among male speakers, so that there are likely to be more dissimilar male speaker pairs

## Additional Analysis/Breakdown of Results [3/3]

---

- ▶ For KL divergence measure, examine the 1%, 3%, 5%, 10%, and 20% most and least similar speaker pairs
- ▶ 150 speakers overall, 87 female and 63 male; 1815 same-sex impostor speaker pairs (with trials in SRE08 short2-short3 task)
  - ▶ Tendency to find 2 types of speakers: those who frequently appear as members of difficult-to-distinguish pairs, and those who occur frequently as members of easy-to-distinguish speaker pairs
  - ▶ 15 speakers (1 male, 14 female) who never appear in the most-similar groups, and 24 speakers (10 male, 14 female) who never appear in the most-dissimilar groups
  - ▶ Supports existence of “wolves” and “lambs”



# Conclusions

---

- ▶ It is possible to predict which speaker pairs will be difficult for a typical speaker recognition system to distinguish
- ▶ Among features considered here, the Euclidean distance between vectors of the mean first, second, and third formant frequencies produces the largest performance difference (on average) for similar and dissimilar speaker pairs
- ▶ Best measure proved to be the approximated KL divergence between speaker-specific GMMs
- ▶ Typically more successful at identifying dissimilar speaker pairs
- ▶ Can provide potentially useful information about a speaker's tendency to be similar or dissimilar to other speakers

# Future Work

---

- ▶ Test combinations of multiple feature-measure as criterion for selecting similar speaker pairs
- ▶ Extend work to find features for selecting target speakers that are difficult for the system to correctly recognize
- ▶ Further investigations into the lack of consistency in how different systems behave for the same set of speakers
  - ▶ Potential trends in behavior across classes or types of systems

# Thank you!

---

- ▶ Questions or comments?