Phillip L. De Leon (New Mexico State University, USA)
Michael Pucher (Telecommunications Research Center Vienna, Austria)
Junichi Yamagishi (University of Edinburgh, UK)

Evaluation of the Vulnerability of Speaker Verification to Synthetic Speech

Contents

- Introduction
- Attacking scenario
- Our previous work
- Speaker verification system
- Speech synthesis system
- Experiments and results
- Detection of synthesized speech
  - IFDLL
  - WER
- Conclusions

Introduction: General background

- It is known that text-prompted speaker verification systems have vulnerability to text-to-speech (TTS) systems
- TTS systems assumed so far
  - Unit selection TTS systems + GMM-based voice conversion
    - Any utterances can be synthesized from only text inputs
    - Output waveforms of the synthesizer can be transformed into a specific targeted person's voice using the voice conversion

- TTS systems in our talk
  - HMM-based TTS systems + speaker adaptation (e.g. MLLR)
    - Speaker adaptation can transform speaker-independent HMMs into the targeted person's model
    - Any utterances can be synthesized from the adapted models
  - This problem was first reported by Masuko et al. 10 years ago

Introduction: Why do we revisit?

- Why do we revisit this issue?
  - The performance of the HMM-based TTS was drastically improved
  - The quality of HMM-based speech synthesis is now comparable with unit selection and its intelligibility outperforms unit selection
  - Enhanced speaker adaptation techniques for TTS
    - Unsupervised adaptation
      - We don’t need to provide labels for adaptation data
      - Two or multi-pass approaches similar to ASR
    - Robust adaptation
      - Noisy data can be used for the adaptation

- It is now possible to automatically create targeted speakers’ TTS voices from any accessible audios which attackers can find.
  - e.g. Audio files available on the web

Attacking scenarios to be assumed

- Speech data is acquired from broadcast, podcasts, lectures, telephone
- Using the acquired speech data, adapt HMM-based TTS systems in advance
- Using the adapted models, synthesize speech for verifications
- Actual synthetic speech samples created in this scenario
  - George W Bush podcast:
    - Synthetic speech samples generated from HMMs adapted using speech data found on his podcasts
    - Sample
    - Real-time demo [web]
  - Queen Elizabeth II's podcast
    - Synthetic speech samples generated from HMMs adapted using speech data found on her podcasts
    - Sample

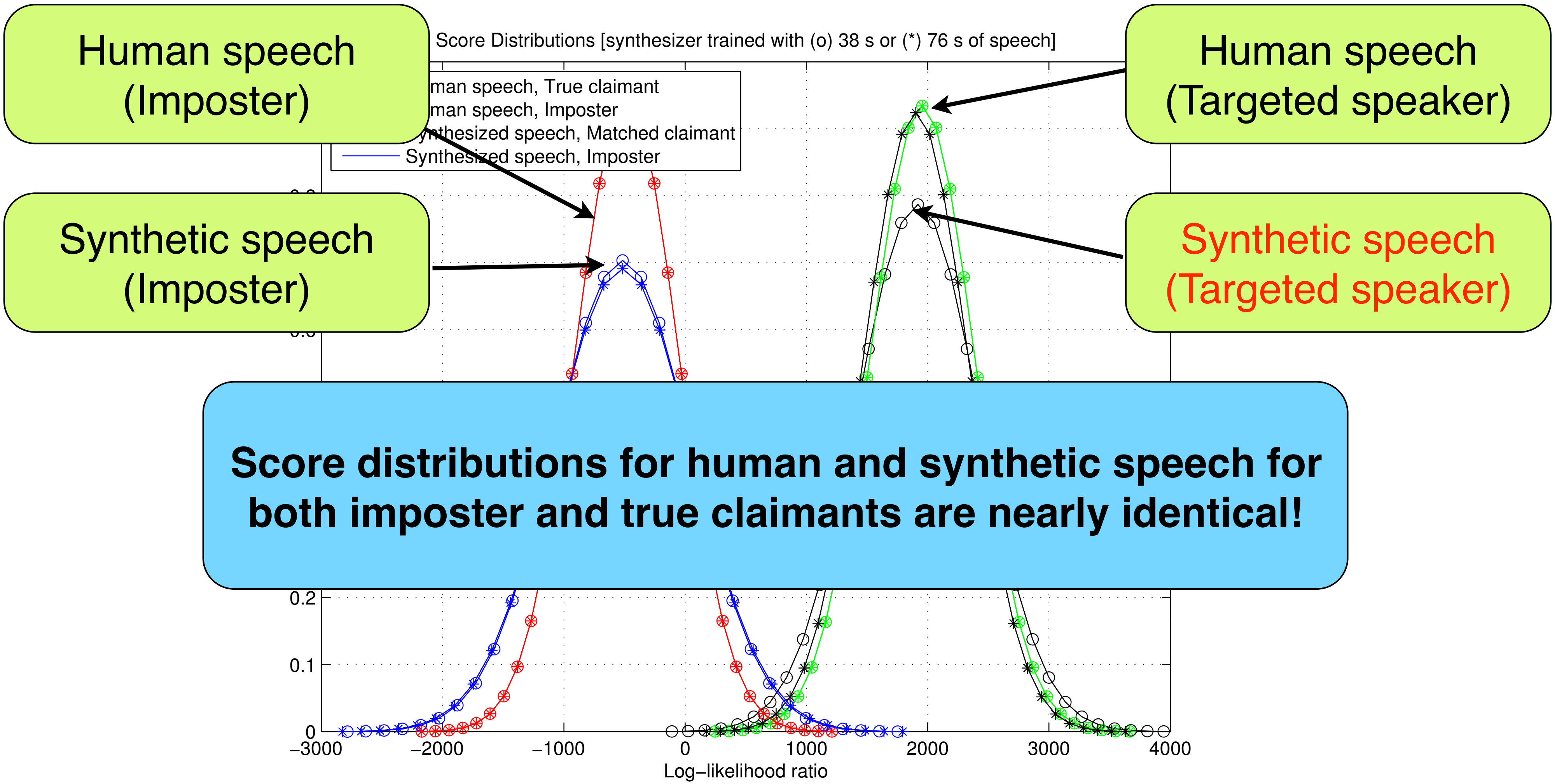
Our previous experiments (ICASSP 2010)

- Speech synthesis databases
  - Perfectly clean read speech
  - Only 10 German speakers

- Two SV systems tested
  - The standard GMM-UBM method
  - Gaussian super-vectors with SVM [C. Longworth and M. Gales 2009]
  - with score normalisation, feature warping/normalisation etc.

- No significant differences from attacker points of view
  - In most of the cases, the SV system will accept a claim from a synthetic voice
  - Report only the GMM-UBM method in this talk

Our previous experiments (ICASSP 2010)



Score distributions for human and synthetic speech for both imposter and true claimants are nearly identical!

Problems of our previous experiments

- The number of speakers is too small
- Perfectly clean speech conditions (In our attacking scenarios, speech data acquired is assumed to be not perfectly clean)

What's new

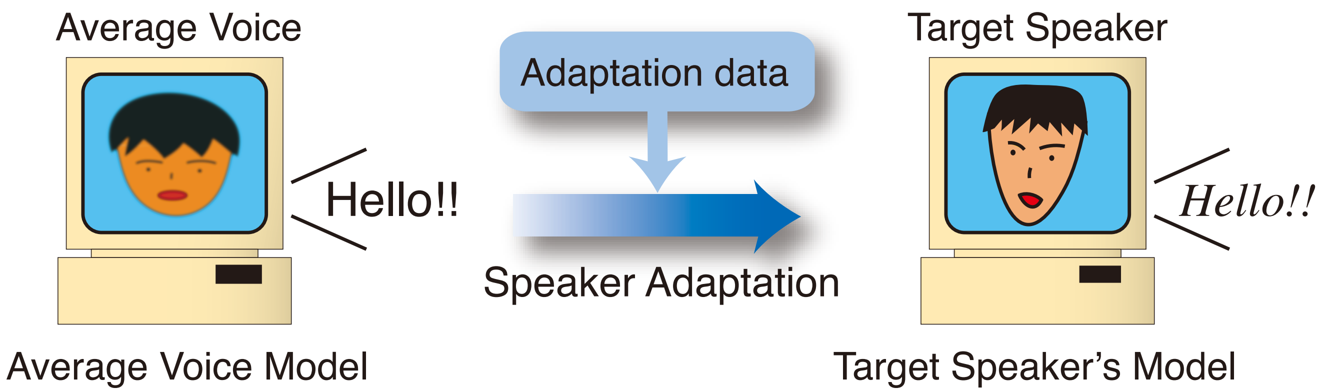
- More speakers: 300 speakers!
  - WSJ corpora SI284 set
  - Not perfectly clean / office environments
  - More realistic conditions

- Report the accuracy of the conventional method to detect synthetic speech in SV systems
  - Satoh et al. reported a method to detect synthetic speech in SV systems in 2001
  - However, the quality of synthetic speech becomes much better than 2001
  - Re-evaluate the method to confirm the problem of imposture using speaker-adaptive HMM-based synthetic speech

GMM-UBM speaker verification system

- GMM-UBM
  - 1024 components
- Features
  - 15 MFCC, 15 Δ-MFCC, log-energy, Δ log-enery
  - Feature warping to improve robustness [J. Pelecanos and Sridharan]
- Adaptation
  - MAP adaptation (mean vector only)
- Performance on the NIST 2002 corpus
  - 330 speakers
  - 12.10% EER
  - Comparable performance with [C. Longworth and M. Gales 2009]

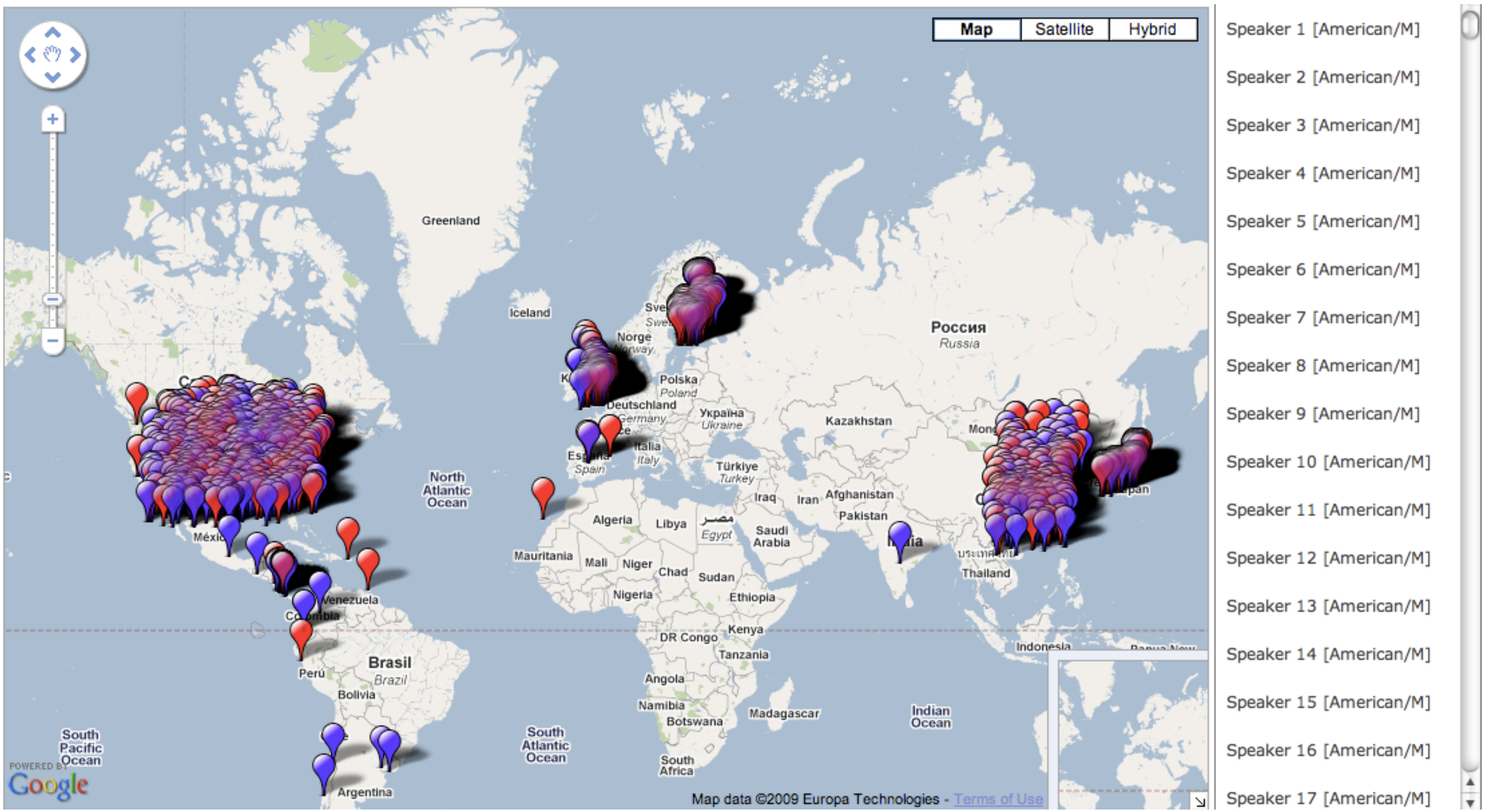
Speaker-adaptive HMM-based speech synthesis



- How to construct the average voice model
  - Speaker adaptive training (SAT) [T. Anastasakos et al. '96]
- How to transform model parameters of the average voice models
  - Speaker adaptation techniques for HMMs
    - Maximum likelihood linear regression (MLLR) [C. Leggetter et al '95]
    - Structural MAP estimation of CMLLR [J. Yamagishi et al '05]
- How to generate acoustic parameters and synthesize speech
  - Maximum-likelihood parameter generation algorithm [K. Tokuda '95]
  - STRAIGHT vocoder [H. Kawahara '2002]

Rapid voice building

- We can rapidly create TTS voices from 3 mins of speech data only
- Currently 2000 voices are created from various sources

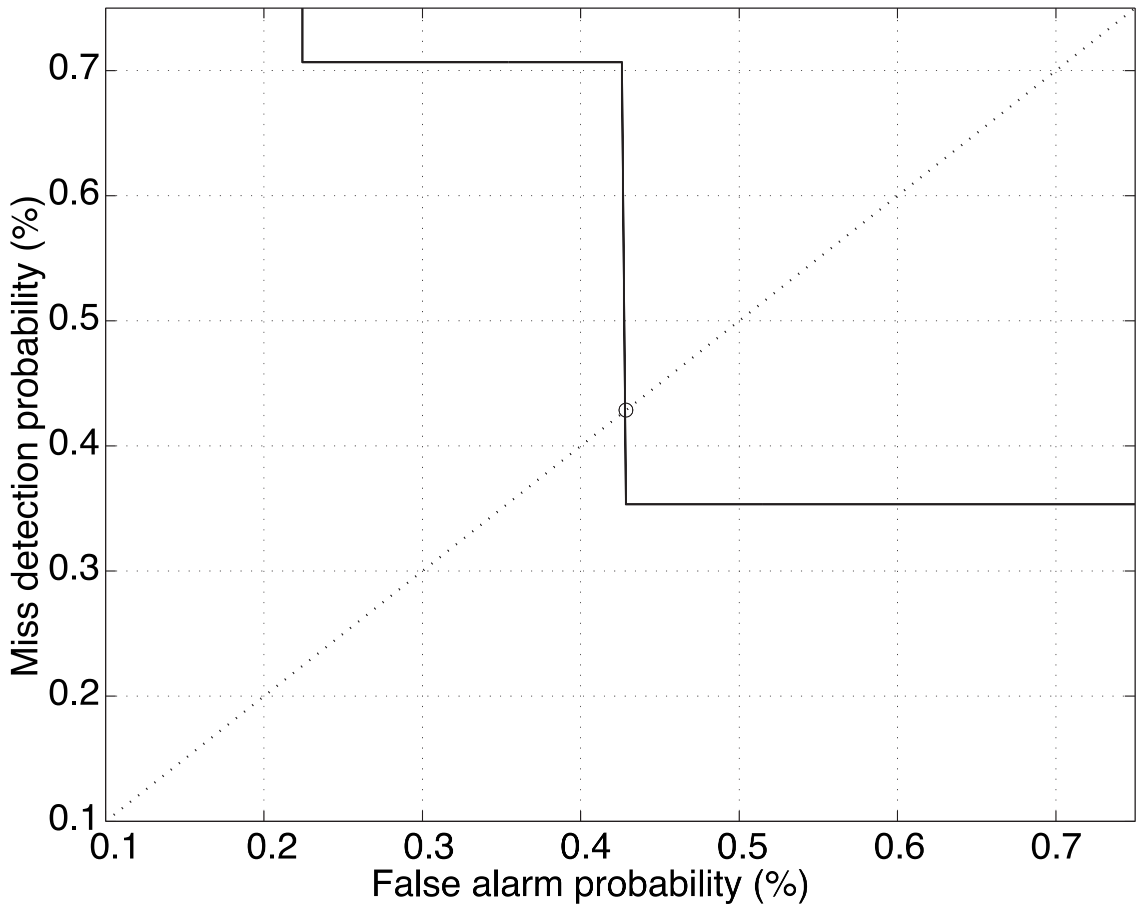


Experiments – Data

- Our scenario is not building TTS systems on speaker verification databases, which are normally narrow band with noises
- Wall street journal corpora (WSJ0 and WSJ1)
  - 283 speakers (included in SI-284 set)
  - Divide the SI-284 speaker material into 3 sets, A, B, and C
  - Set A: TTS training data
    - Training of average voice models
    - Speaker adaptation (CMLLR) to individual speakers
  - Set B: SV training data
    - Training of the universal background model
    - Speaker adaptation (MAP) to individual speakers
  - Set C: Test data (30 sec/speaker)
    - Assumed to be speech reading text-prompts used for verifications
- Samples of synthetic speech created

Experiments – Performance of SV systems

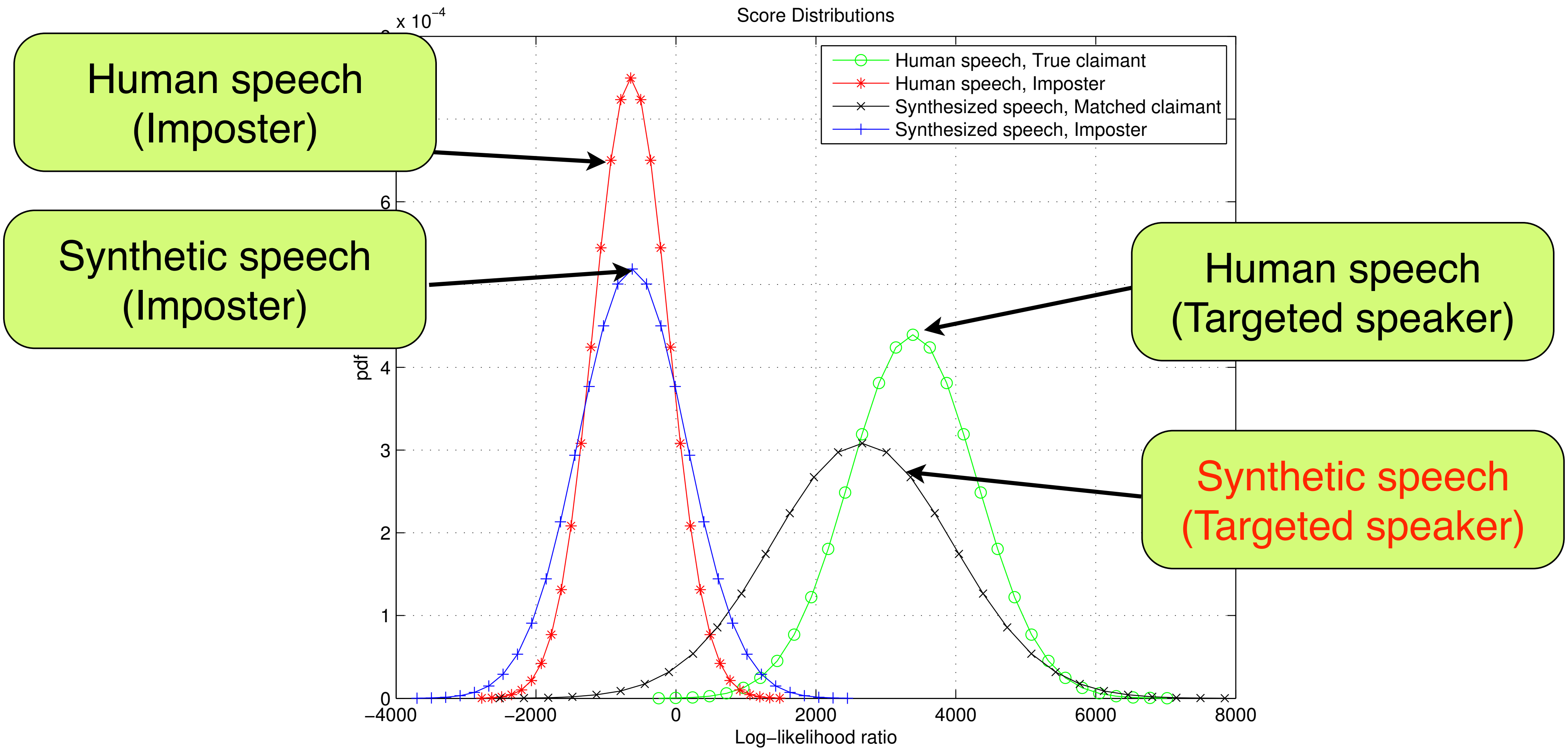
- Decision-error-tradeoff (DET) curve for human speech



- Equal-error-rate is 0.4% (speaker verification of human speech on WSJ corpus is relatively easy)

Experiments – Human speech vs. Synthetic speech

- Score distributions of human speech and synthetic speech



- In matched claimant tests (synthesized voices claim to be their human counterparts), about **90%** of synthetic speech claims was accepted!

Summary so far

- Despite the excellent performance of the SV systems (0.4% EER), the speaker identity of the synthesized speech generated from speaker-adaptive HMM-based speech synthesis is high enough to allow these synthesized voices to pass for true human claimants (90% voices were accepted!).
- Adjustments in decision thresholding or standard score normalisation techniques are unlikely to differentiate between human and synthesized speech

- How can we differentiate them?
  - Detection methods used in the conventional studies
    - Inter-frame differences of log likelihood (IFDLL) [Satoh et al. 2001]
    - ASR word error rates
  - Are they still secure for the latest HMM-based speech synthesis?

Average inter-frame difference of log-likelihood

T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. Eurospeech*, 2001.

- In 2001 Satoh. et al. reported that
  - the average of the inter-frame difference of log-likelihood (IFDLL)

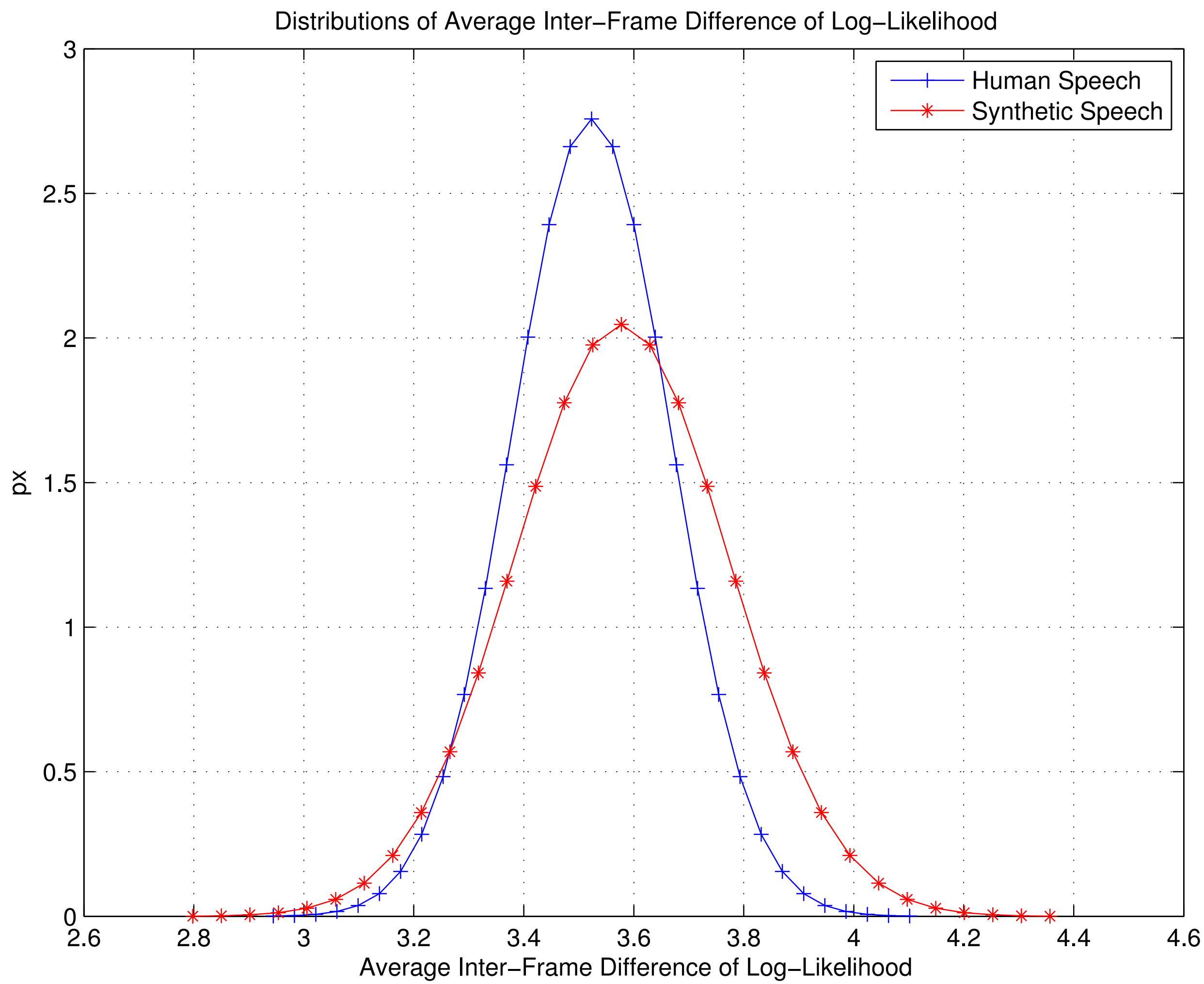
$$\Delta_n = |\log p(\mathbf{x}_n|\lambda_C) - \log p(\mathbf{x}_{n-1}|\lambda_C)|$$

$$\bar{\Delta} = \frac{1}{N}\sum_{n=1}^{N}\Delta_n$$

  - can be used to detect synthetic speech because
    - Synthetic speech generated from HMMs tended to have over-smoothed trajectories (smaller average IFDLL) at that time
    - Synthetic speech using unit-selection tends to have 'jump' at bad concatenation points (larger average IFDLL)

- However the current HMM-based speech synthesis includes global time variation models [Toda et al. 2005], which can avoid the over-smoothing of trajectories.

Results for the average IFDLL



With state-of-the-art HMM-based synthesis this measure is **no longer robust enough** for detecting synthetic speech!

ASR word error rates (WER)

- In the speech perception field, synthetic speech generated using unit selection is known to have less intelligibility than human speech
- It might be possible to see the intelligibility of speech via WERs of ASR

TABLE II
SPEECH RECOGNITION WERS AND SERS IN % FOR WSJ CORPUS (283 SPEAKERS).

Dataset	Grammar3	Grammar4
Human speech	9.55 / 10.79	13.91 / 33.24
Synthetic (73-1620 sec.)	3.05 / 4.85	5.50 / 22.51

- Two grammars tested
- However, synthetic speech was found to have better WERs than human speech on both grammars, even if the adaptation data is 73 sec of speech data
- Not ideal to utilise the WERs of ASR to detect synthetic speech

Conclusions and future work

- Despite the excellent performance of the SV systems, the speaker identity of the synthesized speech generated from speaker-adaptive HMM-based speech synthesis is high enough to allow these synthesized voices to pass for true human claimants.
- This implies that speech data available from e.g. podcasts can be used for imposture in SV systems
- The conventional method using IFDLL to detect synthetic speech is no longer robust enough

- We need to develop new features or strategies to discriminate them!

Q & A