Speaker De-identification using Diphone Recognition and Speech Synthesis

Tadej Justin[1], Vitomir Štruc[1], **Simon Dobrišek**[1], Boštjan Vesnicer[2], Ivo Ipšić[3] and France Mihelič[1]

[1] Faculty of Electrical Engineering, University of Ljubljana, Slovenia

[2] Alpineon Ltd, Ljubljana, Slovenia

[3] Department of informatics, University of Rijeka, Rijeka, Croatia

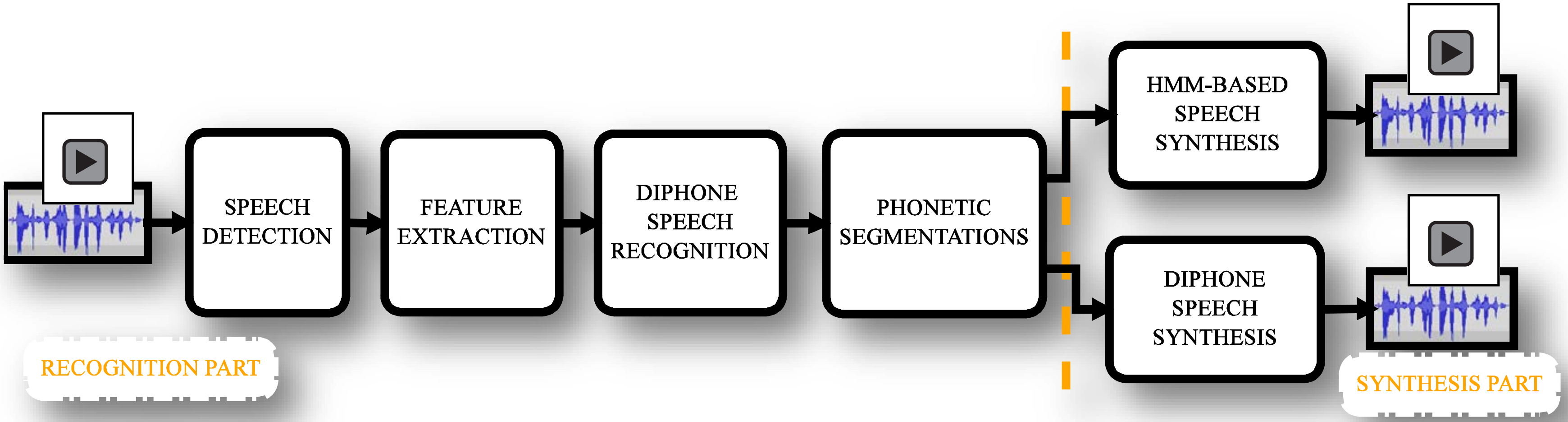
PRESENTATION OVERVIEW

- DROPSY
- System evaluation setup and results
  i. Intelligibility assessment
  ii. De-identification efficacy
- Shortcomings, improvements and further experiments
- Conclusions

DROPSY - Diphone Recognition and Speech Synthesis System

- Speaker de-identification system based on diphone recognition and speech synthesis was developed.
- It is different from other existing techniques that commonly belong to one of the two following groups:
  i. the group of voice-degradation approaches, or
  ii. the group of voice-conversion approaches.

DROPSY



- Speech (phone) recognition module:

  Context-dependent HMM-based bi-diphone acoustical models and a phonetic bigram language model.

- Speech synthesis modules:

  HMM-based or PSOLA-based synthesizers built from the recordings of the two different target speakers

System evaluation and results

i. Intelligibility assessment

Is the de-identified speech still intelligible?

ii. De-identification efficacy

Can I recognize the original speaker?

Intelligibility assessment

- 28 test sentences from the GOPOLIS database:

  7 male and 7 female speakers uttering
  2 different sentences (with 5-8 words).

- 56 (2x28) de-identified speech recordings:

  using two different speech synthesizers.

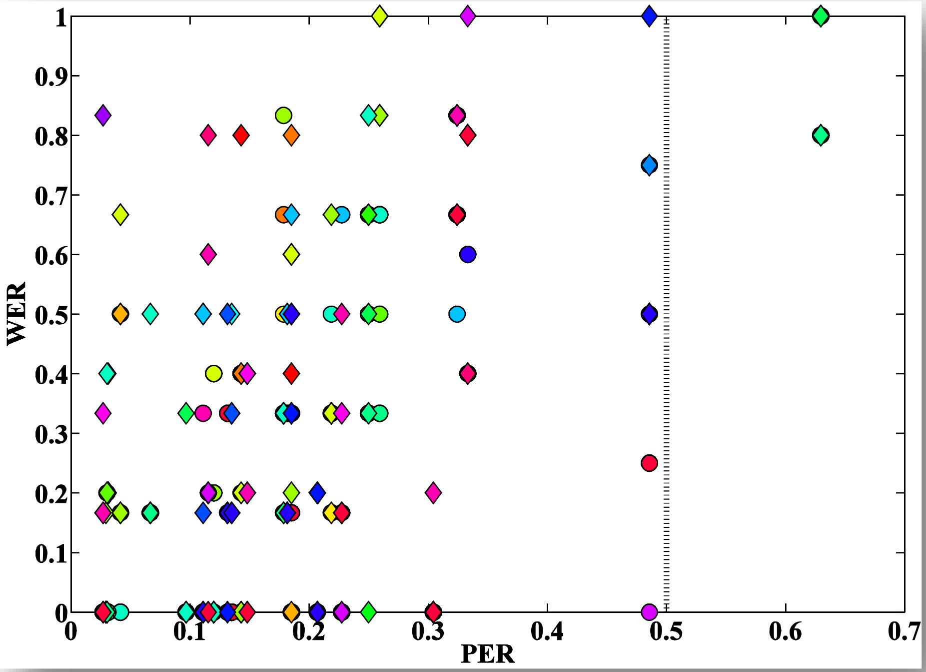
- 26 evaluators (13 males and 13 females) transcribed the de-identified speech recordings.

Intelligibility assessment

- Each evaluator transcribed 14 (2x7) randomly selected de-identified recordings of different test speakers.
- Each evaluator listened to each sentence only once.
- A total of **364** (26*x*14) transcriptions were obtained.
- Word error rates (WER) of the manual test transcriptions were analyzed.
- Phone error rates (PER) of the phone recognizer were obtained from the automatic phone transcriptions.

Intelligibility assessment

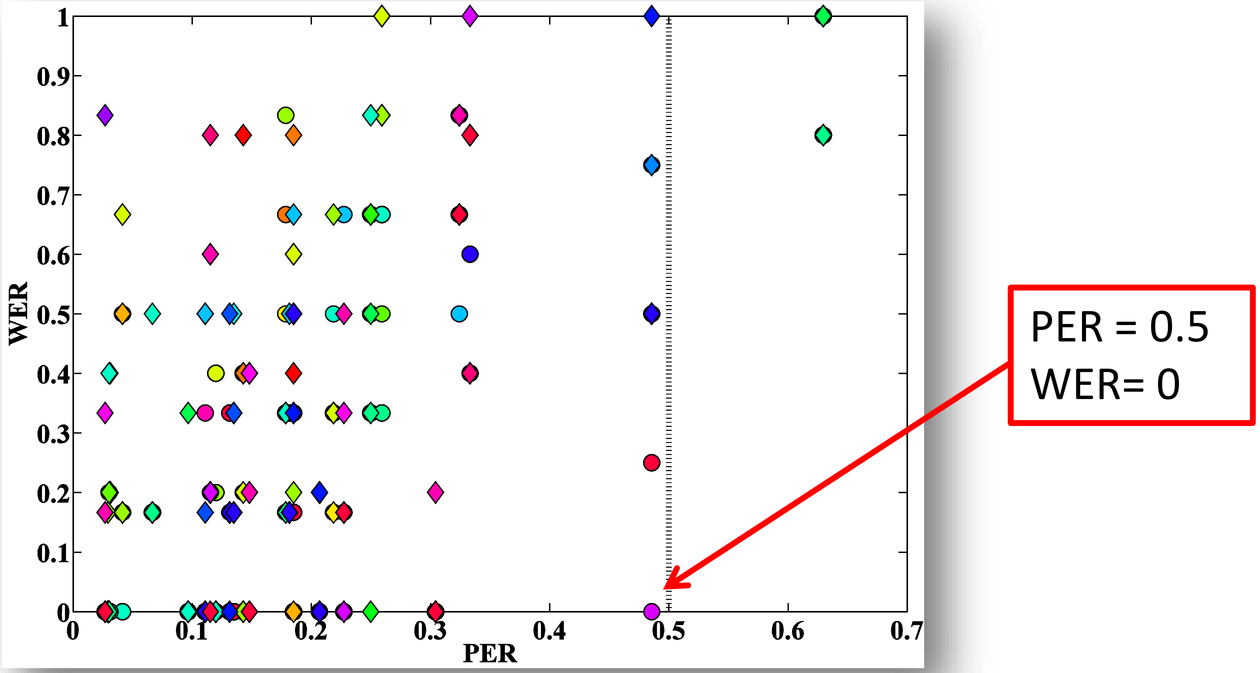
- Word error rates of the listening tests were compared to the phone error rates of the phone recognizer.



- Points on the vertical lines match the transcriptions of different evaluators of the same test sentence.

Intelligibility assessment

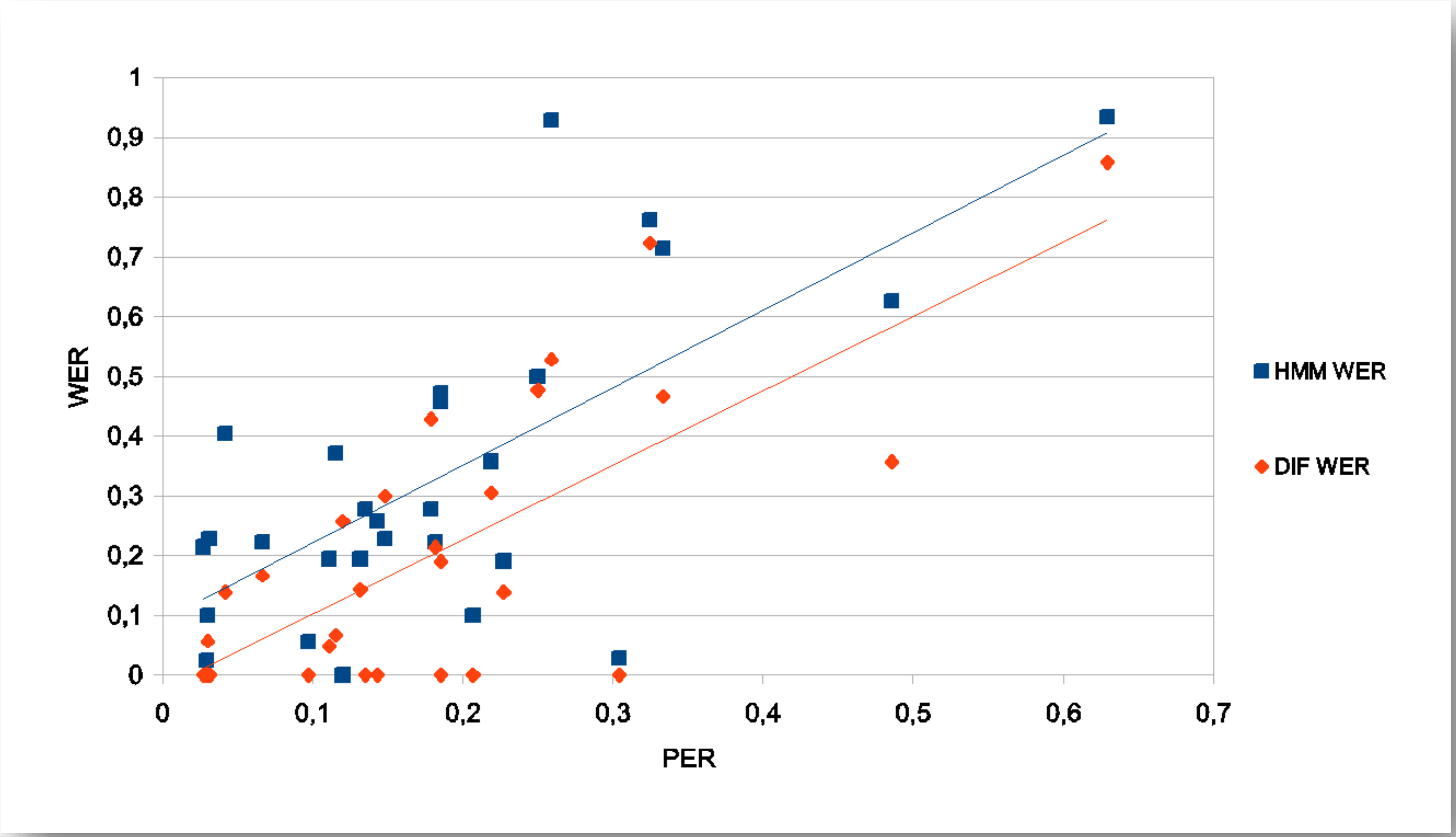
- Word error rates of the listening tests were compared to the phone error rates of the phone recognizer.



- Points on the vertical lines match the transcriptions of different evaluators of the same test sentence.

Intelligibility assessment

Observing average WER in relation to the PER for the two different speech synthesizers.



Intelligibility assessment

- The average WER and PER for all test utterances, depending on speaker's gender, were observed.

GENDER	WER HMM	WER DIF	PER
female	0,44	0,29	0,23
male	0,23	0,13	0,14

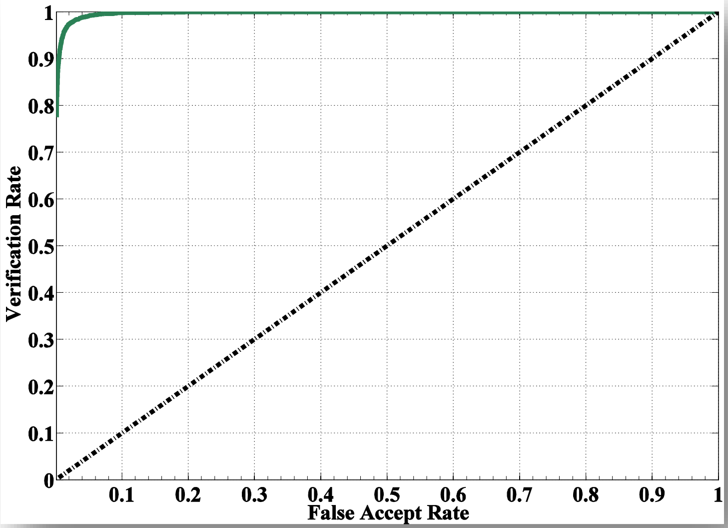
- Intelligibility of the de-identified speech seems to be speaker's gender dependent.

De-identification efficacy

- The use of an automatic state-of-the-art text-independent i-vector-based speaker recognition system.
- The same test speaker identities were used as in the intelligibility test.
- The target speaker recordings were selected from our test database that was not used for building the system.
- Approx. 12 seconds long utterances were used for speaker recognition tests.

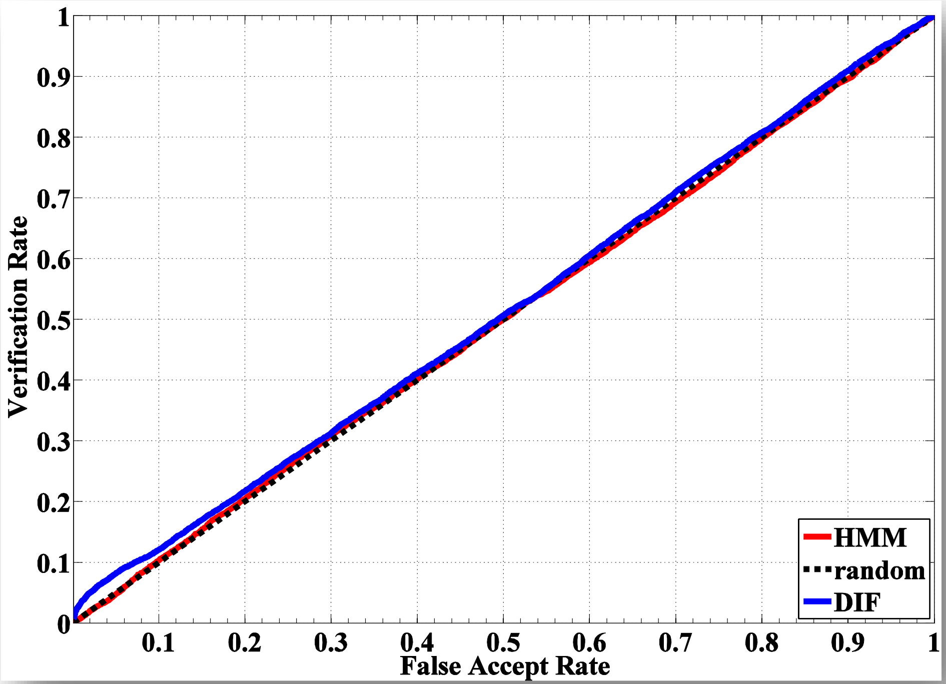
Baseline performance of speaker recognition system

- 8,832 genuine verification attempts and 138,240 impostor verification attempts were conducted using the original (non-de-identified) speech recordings.
- The system achieves a TAR of 77.5% at 0.1% FAR and an EER of 2.36%.



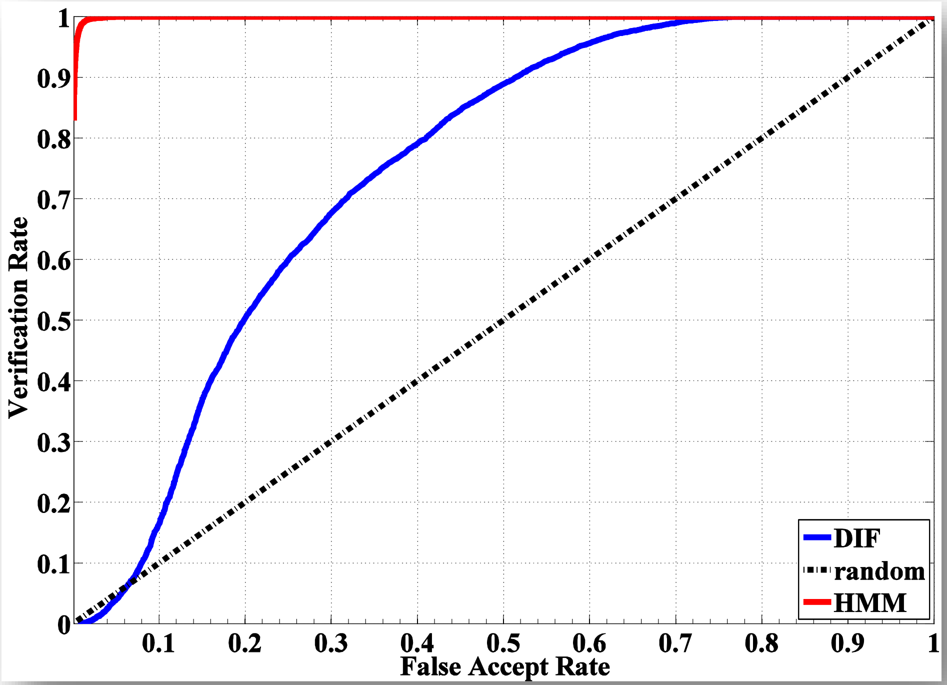
Efficiency of the de-identification procedure

- Speakers were enrolled with the natural speech recordings from our test database.
- Test data includes only recordings of the de-identified speech.



Efficiency of the de-identification procedure

- In the second experiment, we tested the de-identified recordings of the two speakers that were used for building the two speech synthesizers.



Speaker 1

Original

De-identified

Speaker 2

Original

De-identified

Shortcomings, improvements and further experiments

- Performance of the system strongly depends on the accuracy of the phone recognizer.
- The naturalness of the de-identified speech should be improved.
- Different speakers cannot be distinguished from the de-identified speech (voice is always the same).
- The synthesized speech could be transformed to reflect some broader characteristics of the input speaker.

Conclusions

- A relatively novel approach to developing a speaker de-identification system was presented.
- A robust diphone speech recognizer and two different speech synthesizers were combined to build the speaker de-identification system.
- Intelligibility of the de-identified speech was assessed using human evaluators and its efficacy evaluated using a state-of-the-art speaker recognition system.
- The proposed system does not require a full-fledged error-free speech recognition system.

Thank you for your attention!

Questions?