Application of speaker adaptation methods used in speech synthesis for speaker de-identification

Miran Pobar, Department of Informatics, University of Rijeka, Croatia

COST ACTION IC 1206 Meeting, Mataró, Spain, November 25-26 2013

Outline
• Introduction
• Motivation
• Previous research
• The idea/planned work
• Conclusions and future work

Introduction

• Speaker de-identification: natural and intelligible sounding voice, that doesn’t reveal the speaker’s identity [Jin et al., 2009.]
  – Not concerned with identification based on message content
• Applications
  – Call centers, anonymous help lines...

Motivation

• Motivation: application of VT for de-identification, without the need for previous enrollment of new speakers.
• Previous research by [Jin et al. 2009.]: Voice Converging: Speaker De-Identification by Voice Transformation
  – Application of Voice Transformation to speaker de-identification
    • Goal: Voice sounds natural and intelligible, but doesn’t reveal the speaker’s identity
    • Results show successful de-identification while preserving intelligibility
    • Speaker enrollment phase limits applicability

Previous research

• Speaker adaptation in speech synthesis, esp. bilingual speech synthesis
  – A Bilingual HMM-Based Speech Synthesis System for Closely Related Languages [Justin et al., 2012]
  – A Comparison of Two Approaches to Bilingual HMM-Based Speech Synthesis. [Pobar et al. 2013]
  – The basic concepts were already introduced in [Latorre et al., 2006], [Wu et al., 2009], [Liang et al., 2008]

Bilingual speech synthesis

• Bilingual speech system: synthesizing speech in two languages with the same speaker identity.
• Quite a large number of languages with small database resources
  – Hard to recruit speakers fluent in both languages
  – Building the synthetic voice in target language, even though data in the target language does not exist for the target speaker
  – Data from multiple speakers, adaptation to target speaker
• Phoneme mapping approach
  – Data from both languages treated as monolingual, labeled with a joined phoneme set
  – Adapted to the target speaker using the CMNAD adaptation
• State mapping approach
  – Speaker independent models trained separately for both languages
  – Mapping between states derived automatically from minimum KL divergence between states
  – Mapping used to apply transforms calculated for one language to the other language
• How to automatize finding the phoneme/state similarities/differences between spoken data in similar languages?
• Similar method as in de-identification: transformation of one voice to the target voice
Phoneme mapping approach

- Truly bilingual model – speaker independent bilingual voice
  – average voice (training data from multiple speakers in both languages)
  – Speaker adaptive training [Yamagishi et al., 2003.]
- Obtained voice used as basis for adaptation to the target speaker identity
  – CMLLR adaptation of average voice models to the target speaker
- Data labeled with a common phoneme set
  – Conducted with help from a phonetic expert, who has knowledge about both languages.
  – Slovene and Croatian databases already had phonetic dictionaries closely related to SAMPA phonetic alphabets of respective languages, we decided to map the phonemes with a help of SAMPA alphabet.

Phoneme mapping approach

- Database preparation

State mapping approach

- Train two Average Voice models in both source and target languages.
- Target speaker: speaks one language, we want synthesis in both
  – Intra-lingual CMLLR adaptation for the target speaker’s language
  – Application of transforms trained in one language to the other via state mapping
- Establish state mappings between the source state model and the target state models, i.e. find a nearest state model in target language for each state model in source language.
- Train the transforms for the Average Voice model in target language using the adaptation data.
- Adapt each state model in source language using the transform attached to the mapped state model in target language.
- [Wu et al., 2009]
Summary of results

- Results similar for both approaches in intra-lingual situation
- Asymmetric results in cross-lingual situation
- The results suggest that although there are differences in the two approaches of phoneme mapping, the target speaker itself may play important role in the performance of bilingual systems.
- The duration model in cases of S1* for each phoneme is also shared between the trained model, but that is not the case in S2*.

De-identification using VT

- [Jin et al. 2009.]
- Various strategies tested on GMM and Phonetic speaker identification systems
  - Standard Voice Transformation
  - Double Voice Transformation - composition of de-duration + standard VT
  - Transponder VT – linear interpolation of MCEPs x surre + (1-f) x VT
- Results from 92% for GMM and 42% for Phonetic in case of standard voice transformation to 100% for GMM and 87.5% for Phonetic in case of transponder VT
- Better results on large database (95 male and 102 female speakers)
- Human listening tests: intelligibility, quality of de-identification, comparison with naive pitch shifting approach
Speaker de-identification via VT

- [Jin et al. 2009.]
- Voice transformation strategies can be used to de-identify speech
- Standard VT performed well with GMM, but poor with Phonetic
- Investigated different VT strategies: Transsterpolation worked best
- Human tests: de-identified voices are understandable while providing securely transmitted content
- VT targets frame level and some prosodic characteristics BUT
  - SID possible on higher-level aspects, like prosody, word choice
- Secure ways to re-identify a (selection of) speaker(s)
- Need for enrollment of new speakers limits applications

Planned setup

- Database: The multilingual COST2788N corpus (BCN)
  - COST2788N "Audio indexing of Broadcast News files in different languages"
  - complete news shows that were broadcasted by different TV stations in different countries.
  - 16 TV stations, 11 languages (Basque, Croatian, Czech, Dutch, Galician, Greek, Hungarian, Portuguese, Slovakian, Slovenian (2 sets), Spanish)
  - Audio and video: the audio is (16 kHz, 16 bit, wave format) and the video (only 6 data sets also come with video) is mostly in Real Media Video.
  - "The aim is more to provide data that can be used for testing, model adaptation and for research on front-end processing algorithms such as speech/nonspeech and speaker turn segmentation, background classification, speaker clustering, etc. that are supposed to be only weakly language dependent."

Conclusions

- De-identification of speakers using VT
- Suggested improvements increase possibilities for application:
  - Goal: eliminate the need for enrollment of new speakers to be de-identified, which significantly increases application possibilities for de-identification.
  - A pool of transformations, apply transformation according to speaker identification results
  - Expanding set of transformation using data collected from new speakers

Application of speaker adaptation methods used in speech synthesis for speaker de-identification

Miran Pobar, Department of Informatics, University of Rijeka, Croatia

COST ACTION IC 1206 Meeting, Mataró, Spain, November 25-26 2013
References


