New approach to polyglot synthesis: how to speak any language with anyone’s voice

J. Latorre, K. Iwano, S. Furui
Furui Laboratories
Dept. of Computer Science
Main goal of this research

- Create a synthesizer that can speak multiple language with the voice of any person, regardless of the language actually spoken by that person.

- Why? An ever growing number of people use 2 or more languages every day.
  - Bilingual countries: China, India, Pakistan, Belgium, Spain, Paraguay, most African countries, most ex-soviet countries,…
  - 47 million people in USA (18% of the population) speak at home a language other than English. (Census 2000)

=> People who need to speak several languages will expect their computers to do it too.
For which applications is useful a polyglot synthesizer?

- Applications where two or more languages are mixed and a voice switch is not appropriate
- Correct synthesis mix-lingual texts.
  - SOHO is Small Office Home Office的缩写，亦即“小型的、家庭的办公室”的含义。
- Devices that have to be adapted to work in different languages (e.g. speech-to-speech translators, car-navigation systems)
- Help to preserve endangered languages by reducing the development costs
Previous approaches

- Polyglot speaker database
  [Traber et al. 1999]
  - **Advantages**
    - Unit selection speech quality
  - **Disadvantages**
    - Difficult to find polyglot voice talent
    - Hardly expandable

- Phone-mapping [Campbell 2001]
  - **Advantages**
    - Easy and universal
  - **Disadvantages**
    - Too strong foreign accent reduces the understandability
    - Degraded quality in concatenative synthesis
Our approach

• Voice identity depends on anatomical factors.
  ⇒ the average voice of any language should sound more or less the same.

• IDEA ⇒ By mixing data from several speakers in several languages, it should be possible to create an “statistical” polyglot speaker!
HMM-based speaker adaptable polyglot synthesizer

HMM TRAINING

Speaker Independent Voice

MLLR ADAPTATION

Speaker Dependent Voice

HMM-BASED SYNTHESIS

Synthetic speech

Adaptation data: Speaker S in lang. x
x ∈ {A, B, C}

Text in Lang. A

Text in Lang. B

Text in Lang. C
Advantages of this approach

- No real polyglot speaker is required, therefore
  - it can be expanded to any new language.

- No phone mapping is needed, therefore
  - the foreign accent is lower and the intelligibility is better.

- It is based on HMM synthesis, so
  - it can be easily adapted to imitate almost any voice,
  - Small footprint (around 4-6MB for 4 languages).
And disadvantages.

- The audio quality is a telephone-like quality as in any HMM-based synthesizer.
- However,
  - HMM-synthesis can provide better quality than any other synthesis method when the amount of training data is below 50 min [Bennet 2005].
Evaluation (I)

- Compare our method with others based on phone mapping to:
  - Synthesizing the target language with a synthesizer trained in the language of the target speaker.
  - Adapting a synthesizer trained in the target language to the voice of the target speaker.

- We have evaluated the performance of our method according to 3 parameters
  - Perceptual Intelligibility
  - Native accent
  - Similarity to the target speaker
Cross-language synthesis using phone mapping

HMM TRAINING

MLLR ADAPTATION

HMM-BASED SYNTHESIS

Synthetic speech
Cross-language speaker adaptation using phone mapping

- Text in Lang. A
- Adaptation data: Speaker S in lang. B

Phone Mapping

HMM TRAINING

- SI Monoling. HMM Lang B

MLLR ADAPTATION

- SD Monoling. HMM

HMM-BASED SYNTHESIS

Synthetic speech
Evaluation (II)

- We have considered three different scenarios:
  - Cross-language synthesis: The language spoken by the target speaker and the language to be synthesized are different but included in the training data of the polyglot model.
  - Synthesis of extrinsic languages: The language to be synthesized is not included in the training data.
  - Direct synthesis: The language spoken by the target speaker and the language to be synthesized are the same (and included in the training data)
Experimental conditions

- Evaluation method: Subjective evaluation in a 5 points MOS scale.
- Evaluation Language: Spanish and Japanese.
- Subjects: 6 native speaker for each evaluated language.
- Languages used to train the synthesizers: Different combinations of Russian, French, German, Spanish and Japanese.
- Models adapted to two target voice for each language included in the mixture: 66 SD models.
- Test sentences: 18 different sentence synthesized by each SD acoustic models.
System Details

- **Speech Data:**
  - Globalphone, general purpose databases
  - Training data: 10 speakers for each fully included language with ~10 minutes of data for each speaker
  - Adaptation data: 10 minutes of data for each target voice.

- **Models:**
  - Triphone HMMs, 3 states, 1 Gaussian.
  - 25 MELC and their delta from a 16ms window.
  - Single root tree clustering.
  - The models were adapted to the target voices with supervised MLLR using 4 adaptation classes.

- Original prosody (f0 and duration) from the audio version of the test texts.
Cross-language synthesis scenario

The language spoken by the target speaker and the language to be synthesized are different.

![Cross-language synthesis chart]

**Perceptual Intelligibility**
- Monolingual phone mapping synth.
- Monolingual phone mapping adapt.
- Polyglot cross
- Vocoder

**Native Accent**
- Monolingual phone mapping synth.
- Monolingual phone mapping adapt.
- Polyglot cross
- Vocoder

**Similarity to target spk.**
- Monolingual phone mapping synth.
- Monolingual phone mapping adapt.
- Polyglot cross
- Vocoder
Synthesis of extrinsic languages

- To create a speech synthesizer for a new language is a very expensive task, only profitable for a dozen or so languages.

- For minority language a possible solution is to use speech resources which are available from a phonetically similar language.
Synthesis of extrinsic languages with a polyglot synthesizer

HMM TRAINING

MLLR ADAPTATION

HMM-BASED SYNTHESIS

Synthetic speech
Synthesis of extrinsic languages scenario

The language to be synthesized is not included in the training data.
Perceptual intelligibility vs acoustic distance

- Spanish Coef = 0.97
- Total Coef = 0.98
- Japanese Coef = 0.99
Direct synthesis scenario

The language spoken by the target speaker and the language to be synthesized are the same.
Demo cross-language

- El pasado lunes fue el día de los trabajadores en Estados Unidos 新党 準備会実行委員長を務める tradicionalmente se señala 小沢代表幹事の周辺は

Cross-language synthesis

- MOS
  - Perceptual Intelligibility
  - Native Accent
  - Similarity to target spk
Demo extrinsic languages

- La consigna de los seguidores del nazismo, se llamaba colaboracionismo, esto es el apoyo activo a una potencia de ocupación enemiga.
Conclusions

- It is possible to create a polyglot synthesis by mixing corpora of different languages.

- The performance of a polyglot synthesizer is better than methods based on phone-mapping when
  A) the language of the speaker is different than the language that is synthesized (Cross-language synthesis).
  B) there is no available speech data from language to be synthesized, (synthesis of extrinsic languages).

- In the normal case, the performance of the polyglot synthesizer is equivalent to that of a standard monolingual synthesizer in that language.
Next steps

A) Improve the audio quality: GV, HNM, trajectory HMM, etc.
B) Improve the speaker adaptation: SAT, SMLLR
C) Test the amount of speech data needed to synthesize a new language with the same performance as the languages previously included.

- Check which approach can be applied to the prosody.
Thank you very much for your attention