論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	New approach to cross-language synthesis
著者(和文)	岩野 公司, 古井 貞熙
Authors(English)	Javier Latorre, Koji Iwano, Sadaoki Furui
出典(和文)	日本音響学会2005年秋季講演論文集, Vol. 〔3-6-12〕, No. , pp. 351-352
Citation(English)	, Vol. 〔3-6-12〕, No. , pp. 351-352
発行日 / Pub. date	2005, 9

New Approach to Cross-language Synthesis

◎ Javier Latorre, Koji Iwano, Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

The development of speech technology in a new language is an expensive task that requires the acquisition of speech and text data together with a good amount of man-power and knowledge about that language. A method to reduce part of these costs is to reutilize the resources already collected for other languages. For this, the sounds of the target language have to be modeled by those of the available one, for example by means of a phone mapping [1]. But mapping introduces an error, which depends on the accuracy of the mapping, i.e. the similarity between original and mapped sounds. The larger the error, the lower the performance we can obtain.

2 Phone mapping

If we use the data of only one source language, the mapping error has a lower limit given by the phonetic similarity between that language and the target one. A way to overcome this limit is to use data from more than one source language. In this way, the palette of sounds available can be expanded, and the mapping error reduced. In speech recognition this idea can be used directly [2] because usually what we want to have is a speaker-independent recognizer. In speech synthesis however, we cannot synthesize each sound of a word with a different voice. We need to be able to synthesize the sounds of different source languages with the same voice, i.e. we need a polyglot synthesizer .

3 HMM-based polyglot synthesizer

In [3] we proposed a new approach to the polyglot problem based on HMM synthesis. Briefly, our method consists in training a speaker-independent HMM-based synthesizer with the data of different speakers in different languages so that it becomes also language-independent. This synthesizer is then adapted by means of MLLR to a specific target speaker. The adaptation produces a personalized and coherent voice. Using the adapted models, it is possible to synthesize any of the languages included in the training of the polyglot synthesizer with the same voice and quality, independently of the original language of the target speaker. An additional advantage of our polyglot synthesizer over others based on real polyglot speakers is that we can always expand the inventory of sounds by adding a new language to the training data.

4 Experiments

The purpose of this experiment was to compare the performance of synthesizing Japanese by a monolingual synthesizer in a language phonetically very similar to Japanese: Spanish, with three polyglot synthesizers trained with a mixture of Spanish and other two languages, and a polyglot synthesizer trained with Spanish and Japanese data (Sp+Ja).

4.1 Trilingual models

We have trained three speaker-independent HMM synthesizers by combining data from: Spanish, German and French (Sp+Ge+Fr), Spanish, Russian and German (Sp+Ru+Ge) and Spanish, Russian and French (Sp+Ru+Fr). Each model was trained with around 5 hours of speech: 10 speakers for each language and around 10 minutes for each speaker. All the data belong to the Globalphone Corpus.

The models are triphone HMMs with a single Gaussian and left-to-right 3 states without skips. The feature vector consists of the 25 first MFCC and their delta. The models were clustered using the MDL criterion. Each model was adapted using unconstrained MLLR with 4 adaptation classes to the voice of six speakers, two speakers for each language, yielding a total of 18 adapted models.

4.2 Evaluation

The evaluation parameters were the perceptual intelligibility and the level of foreign accent of 18 Japanese texts synthesized by the 18 speaker-adapted trilingual models. By "Foreign Accent" we mean whether the speech sounds as a native Japanese speaker (5 points MOS) or rather as a foreigner who does not speak Japanese (1 point MOS). This gives us an idea about the naturalness of the synthetic speech.

言語横断音声合成にむけた新たなアプローチ
ラトレ・ハビエル，岩野 公司，古井 貞熙 （東工大）

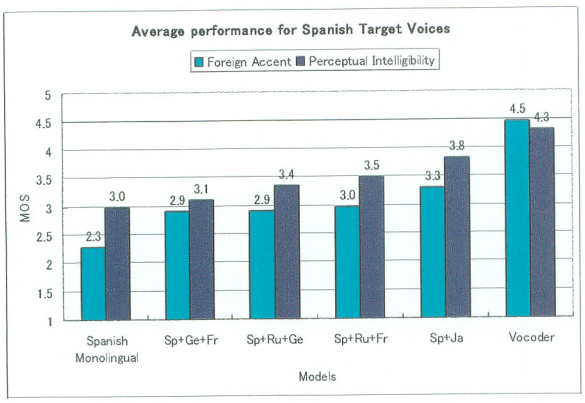

Fig. 1 Performance for Japanese synthesis by Spanish target voices

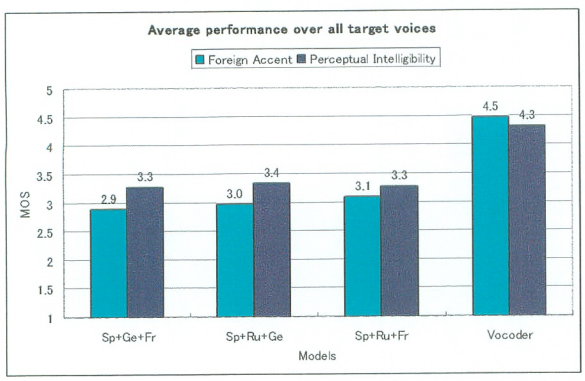

Fig. 2 Average performance over all target voices

We asked 6 native Japanese speakers to evaluate these two parameters in a 5 points MOS scale. Each subject listened to 72 audio files in a random order: 3 files for each adapted models plus the vocoder re-synthesis of the evaluation texts. We have included the vocoder re-synthesis as a reference so that we can establish direct comparisons with previous evaluations

4.3 Prosody

In order to focus only on the phonetic characteristics, we used original prosody extracted from the speech files corresponding to the evaluation texts. To approximate the original prosody to the target speakers, we have shifted the mean f_0 of the test files to the mean f_0 of the target speakers.

5 Results

Figures 1 shows the average performance for Spanish target speakers for each language mixture, and Figure 2 the average performance over all the target voices. The light and dark color bars in each figure show the "foreign accent" and "perceptual intelligibility" respectively. We can see that the perceptual intelligibility of all the polyglot models is better than that of the monolingual model. The average intelligibility over all the target voices was basically the same for the three language mixtures.

With respect to the foreign accent, there is even a greater difference between the monolingual and the polyglot models, to the point that for some target voices the difference in the level of foreign accent of the Sp+Ru+Fr model and the model trained with Japanese data was statistically insignificant.

6 Conclusions

In this paper we have evaluated three HMM-based polyglot synthesizers when synthesizing a language not included in their training data by means of phone mapping. We have shown that the perceptual intelligibility and level of foreign accent of a polyglot synthesizer equal or surpass that of a monolingual synthesizer in a language phonetically very similar to the target one. Furthermore, in some cases the level of foreign accent of the polyglot synthesizer does not differ significantly from the foreign accent of a synthesizer that includes the target language.

7 Future Work

Our next task is trying to go beyond phone mapping. For this we want to consider not just the phone itself but also its context e.g with triphone mapping. Also, we will try to create acoustic models for the unseen sounds not by mapping but by interpolating the models of already existing triphones.

Another item we want to study is to which extent the concept of mapping can be applied to the prosody.

8 Acknowledgements

This work was partially funded by the 21st Century COE-Large-scale Knowledge Resources Program.

References

[1] Campbell, N., "Talking foreign. Concatenative speech synthesis and the language barrier", *Proc. Eurospeech*, pp. 337-340, 2001.

[2] Schultz, T. and Waibel, A., "Experiments on cross-language acoustic modeling", *Proc. Eurospeech*, pp. 2721-2724, 2001.

[3] Latorre, J., Iwano, K. and Furui, S., "Polyglot synthesis using a mixture of monolingual corpora", *Proc. ICASSP*, pp. 1-4, 2005.