論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Toward an HMM-based polyglot synthesizer
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 2004年秋季講演論文集, Vol. , No. 3-2-3, pp. 321-322
Citation(English)	, Vol. , No. 3-2-3, pp. 321-322
発行日 / Pub. date	2004, 9

Toward an HMM-based polyglot synthesizer

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

1 Introduction

The most interesting advantage of speech synthesizers over human beings is probably the ability of speaking various languages. An increasing number of applications such as e-mail reading or car navigation systems require this multi-lingual capacity. Although it is possible to switch to a different output voice when the input language changes, this might not be appropriate in most cases, like for example reading directions of a foreign country in a car navigation system.

The traditional approach toward the polyglot issue has been based whether on a phoneme mapping between the original and the target language [1], or in a recording of multilingual data from a real polyglot speaker [2].

Although the first method can produce acceptable results, especially for phonetically close languages and with a carefully tuned mapping, the resulting speech retains the foreign accent of the original speaker very strongly which can deteriorate the intelligibility.

The second method can offer the quality of unit-selection synthesis, but since it requires a polyglot speaker, it is hardly expandable to more than 4 or 5 languages. Even for only two languages, it may be extremely difficult to find good bilingual speakers for some unusual combinations.

For some applications e.g. voice-to-voice translation systems, it would be desirable to convert the output voice of the synthesizer to the user's voice. In order to achieve this for target speakers that do not speak the target language, some kind of cross-language voice conversion as the one described in [3] has to be applied. This might also require a mapping between the language of the target speaker and the languages of the corpus.

2 Description of the system

Our polyglot synthesis system is based on HMM-synthesis as described in [3]. In our approach we tried to substitute the need of recordings from a real polyglot speaker by a Speaker-, Language-Independent (SLI) central voice created from a set of monolingual speakers. The hypothesis is that such a central voice depends only on anatomical factors, i.e. it is language independent.

In this experiment we tried to synthesize Japanese speech with voice quality of a Spanish target speaker.

2.1 Database

As a first step two phonetically similar languages, (Japanese and Spanish) were combined. We selected these languages for their phonetic similarity, and for the availability of language resources and test subjects.

The corpus that we have used to train our system is GlobalPhone[5]. GlobalPhone is a multilingual corpus that contains recordings of numerous speakers in 16 languages.

The data sampled at 16 KHz were windowed by a 30 ms Blackman window with a 5 ms shift. The feature vector consists of 25 mel-cepstral coefficients and their delta coefficients.

2.2 HMM models

Three speaker independent models were trained: a Japanese monolingual, a Spanish monolingual and a bilingual model. The Spanish model was trained with 104 minutes of data from 10 speakers; the Japanese with 114 minutes from 10 speakers, and the bilingual with the summation of the training data of both monolingual models.

The HMM triphone models are 4 mixtures, 3 states left-to-right models without skips. The transitions between states are modeled by transition matrices.

To avoid an early mixture of the Japanese and Spanish triphones in the bilingual model, a language tag was added to the transcription labels. These language-tagged models were then clustered using one single phonetic decision tree for all the triphones. The inclusion of a question about the language in the decision tree was tested but it did not produce any noticeable difference with those models clustered with phonetic questions only.

2.3 Speaker adaptation

The mean values of the SI models were adapted with supervised MLLR [6] to a Spanish target speaker. The Spanish and the bilingual SI models were adapted directly with the adaptation data of the Spanish target speaker. To adapt the Japanese SI Model, the phonemes of the Spanish transcription of the adaptation data had to be previously mapped onto Japanese phonemes. This mapping was done by rules based on the difference in the articulatory features and our own subjective criteria.

3 Synthesis

To synthesize Japanese speech with the Japanese and bilingual models, the Japanese phonetic labels were directly used as input to the cepstrum generator. For the Spanish model however, these labels had to be previously mapped onto Spanish. This mapping was done by rules, based on the difference between the articulatory features of the phonemes and our own subjective criteria.

A pulse-noise schema has been used to model the source excitation. Since our main focus is the segmental, we decided to use original prosody. We used the duration values of the HMM alignments of the original data and the pitch extracted from the original audio files by ESPS get_f0. This pitch was synchronized with the duration and modified to imitate the tone and pitch range of the target speaker.

4 Evaluation

The parameters that we wanted to evaluate were intelligibility of Japanese synthetic voice, similarity

---

HMM を用いた多言語音声合成の実現に向けての検討

ハビエル　ラトレ，岩野 公司，古井 貞熙 （東工大）

between the adapted synthetic voice and the voice of the Spanish target speaker, and level of "foreign accent".

First, for each model type three clustered models were generated using two different clustering thresholds and the MDL criterion. These clustered models were then adapted to the voice of a Spanish speaker with 1, 4, 16, 64, and 256 adaptation matrices. For each model type we pre-selected the combination of the clustering threshold, i.e. total number of states, and the number of adaptation matrices that produced the HMM model with the best trade off between quality for synthesizing Japanese text and similarity to the original Spanish speaker. Table1 shows the number of adaptation matrices and number of states of the pre-selected speaker adapted models.

Table1 Pre-selected variants of the original model

Model type	N. states	N. Adapt.Matrices
Japanese	400	16
Spanish	1746	4
Bilingual	2265	16

In the case of the bilingual model, around 40% of the states were shared by Japanese and Spanish triphones. In general, increasing the number of adaptation matrix improves the similarity to the original speaker but degrades the synthesis of Japanese.

In addition to the three described methods, we included for comparison purpose the vocoder reconstruction of the original cepstrum with pitch modification and the Spanish diphone synthesizer included in Festival. In order to synthesize Japanese texts with Festival, the Japanese phonemes were transcribed onto Spanish with the same phoneme mapping as for the HMM monolingual Spanish model.

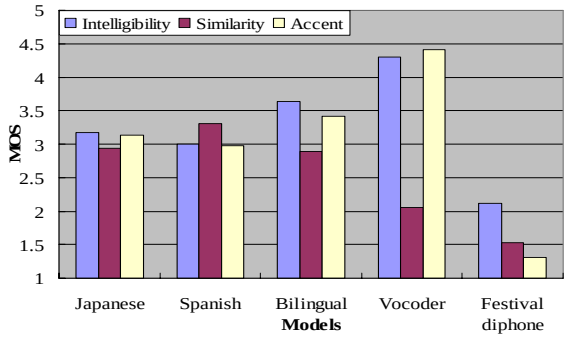

Fig. 1. Results of the MOS test

For the evaluation, 12 Japanesegb texts were synthesized by all the 5 methods. Six native Japanese speakers evaluated the above mentioned parameters for 6 utterances for each model over a 5 points scale. At the beginning of each session, every subject listened to a vocoder reconstruction of a Japanese utterance not included in the test set. This utterance was asked to be considered as the level 5 reference for accent and intelligibility.

To evaluate the similarity between the synthesized Japanese and the original Spanish speaker, a vocoder reconstruction of an utterance of the original speaker was presented before every Japanese utterance.

Fig. 1 shows the results of the evaluation. The intelligibility of the bilingual model is a half point higher than that of the monolingual models. No significant difference was found between the Spanish and Japanese monolingual models. In all cases the Festival Spanish diphone synthesizer was considered worse than any of the other methods. Similar results were found for the "foreign accent" evaluation.

For the similarity, the Spanish monolingual model outperforms, as expected, all the other methods. No significant difference was found between the Japanese monolingual and the bilingual model.

5 Conclusions

A new approach to a polyglot synthesizers using HMM-based synthesis has been proposed. The proposed method uses a combination of monolingual speakers of multiple languages to create a polyglot central voice that can be easily adapted to any target speaker.

The evaluation shows that for cross-language synthesis HMM-based methods outperform diphone concatenation with phoneme mapping. It also shows that the intelligibility of the proposed method outperforms the other methods based on monolingual models and phoneme mapping.

6 Future Work

Our future work is oriented in three directions: increase the number of language for the polyglot model, improve the quality of the synthetic speech and improve the voice adaptation.

We expect to add soon another language phonetically closed to the two already clustered, e.g. Korean or Italian.

To improve the quality and the voice adaptation we are considering applying different clustering and central voice techniques. We also want to try different ways of phoneme mapping based on the phonetic decision tree.

7 Acknowledgements

We would like to thank Drs. A. Black, T. Schultz and T. Toda at CMU, Dr. K. Tokuda at Nitech and Dr. A. Bonafonte at UPC for many helpful discussions.

This work is supported in part by 21st Century COE-LKR Program.

8 References

[1] Nick Campbell, "Talking Foreign. Concatenative Speech Synthesis and the Language Barrier", in *Proc. Eurospeech 2001*, pp. 337-340

[2] C. Traber, B. Pfister et al. "From Multilingual to Polyglot Speech Synthesis", in *Proc. Eurospeech 1999*, pp. 835-838

[3] M. Mashimo, T.Toda, K.Shikano, N.Campbell "Evaluation of Cross-language Voice Conversion based on GMM and STRAIGHT", in *Proc. Eurospeech 2001*, pp. 361-364

[4] T.Masuko, K.Tokuda, T.Kobayashi and S.Imai "Speech synthesis using HMMs with dynamic features" in *Proc. ICASSP 1996*, pp. 389-392

[5] T.Schultz "GlobalPhone: A multilingual Speech and text database developed at Karlruhe University" in *Proc. ICSLP 2002*, pp. 345-348

[6] M.Tamura, T.Masuko, K.Tokuda, T.Kobayashi "Speaker Adaptation for HMM-based speech synthesis system using MLLR" in *Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 273-276