論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Addition of new languages to a polyglot HMM-based 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2005年春季講演論文集, Vol. , No. 1-1-22, pp. 197-198
Citation(English)	, Vol. , No. 1-1-22, pp. 197-198
発行日 / Pub. date	2005, 3

1—1—22

Addition of new languages to a polyglot HMM-based synthesizer

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

1. Introduction

English is nowadays the main language for international communication. However, this does not mean that English is going to become the only world language with exclusion of all the others. Indeed, the globalization process and the economical and demographic changes point rather toward a future where people will have to use two or more languages in their daily life [1]. In such a future, users will demand more and more applications that help them to interact with people that speak a different language, without being need to learn that language. In such context, multilingual capacity is likely to become crucial in the coming years.

For multilingual speech recognition there have been several researches e.g. [2], however for the synthesis of several languages with the same voice individuality there have been until recently only two proposals: based on a corpus from a polyglot speaker [3] or in a phone-mapping across languages [4]. In [5] we proposed a method based on HMM synthesis that combines monolingual corpora from several speakers to create a speaker adaptable polyglot synthesizer. This method could overcome some of the limitations of [3] and [4]. At first, we applied it to two phonetically close languages: Spanish and Japanese. In this paper we present the results of adding to the previous bilingual system a third language, Icelandic, that is phonetically distant from both Spanish and Japanese. We have also evaluated the effects of adding a new language, when the system is adapted to a speaker of an external language not included in the training data.

2. The system

The process of building a multilingual HMM synthesizer is basically the same as for a monolingual one [6]. First, speaker and language independent triphone models are created. To compensate the lack of training data for some triphones, the original HMMs are tied using a phonetic decision tree and retrained. Then, the speaker independent HMMs are adapted with MLLR to the voice of a specific speaker[7]. Finally the adapted HMMs are used to synthesize the output audio.

We use 3-state left to right triphone models without skipping for the first 25 mel-cepstrum coefficients.

In our experiments we use the original prosody of the test samples. Only the mean value and dynamic of the pitch are modified to resemble those corresponding to the target voices.

2.1 Changes respect to the previous system

We have introduced several modifications to our previous system in the labeling of the phonetic sequence and in the parameters of the system.

In our previous experiments, we labeled each phoneme with a language tag. In this way, we prevented an "a-priori" mixing of the phonemes from different languages. The triphone models for each language were trained separately, and clustered according to a single phonetic decision tree for all the phonemes. We did not include any question about the language tag; therefore phonemes with the same articulatory features were always clustered together. However, by training the models separately, their combined values and statistics are different than if they would have been trained together. This affects the general structure of the tree, i.e. which phonemes are tied to which one. In the new system we have eliminated the language tags. We assumed that, if two sounds share the same IPA symbol they are similar enough to be considered as the same sound.

Another modification we have done to the labels is the substitution of some models, like diphthongs and palatalized phonemes, by the combination of two.

Due to the mentioned substitution of the diphthongs models, the minimal length required to model some sounds increases. If the analysis window is too long, the minimal time covered by the sequence of HMMs can even exceed the real time of the sound. To avoid this, we have changed the length of the analysis window from 32ms to 16 ms.

With the inclusion of a third language, the amount of data and processing time increase. To reduce the processing time, we have moved from a 4-mixture model to a single-mixture model.

2.2 Training data

The training data consists of ten speakers for each language speaking approximately 10 minutes each. The adaptation data consist of around 10 minutes from speakers not included in the training data. The Spanish and Japanese data belong to the Globalphone corpus [8]. The Icelandic belong to the Jensson's Corpus, collected by a student at Furui's Lab. It should be noted that none of these corpora were specifically created for speech synthesis.

2.2 Clustering tree

We have created 5 HMMs: one monolingual model for each language (Icelandic, Spanish and Japanese), a bilingual HMM for Spanish and Japanese, and a trilingual one with all three languages. All the models were clustered with a phonetic decision tree and the same threshold for the ML criterion. For the monolingual models, a tree with 3561 leaves for Spanish, 3331 for Japanese and 4213 for Icelandic were produced. The trilingual tree produced a model with 10820 leaves and the bilingual Spanish-Japanese tree produced a model with 6654 leafs.

The level of clustering across languages is difficult to determine, because we have eliminated the language tag. However, we can estimate the level of phonetic coverage for each language as the summation of the frequency of all the phonemes that are shared with at least one other language. This phonetic coverage is 92.5% for Spanish, 69.2% for Japanese and 62.5% for Icelandic. Table 1 shows the phonemes that belong to 1, 2 or 3 languages.

All the HMMs were adapted with MLLR using 4 matrices.

Table 1: Phoneme sharing

All three languages	F, I, j, k, m, n, o, p, s, t, ɰ
Only two languages	B, d, e, g, ŋ, tʃ, z, ð, ɣ, a, l, r, u, h
Only one language	ʔ, ʃ, ɯ, ɑ, dʒ, ɻ, ts, ʎ, θ, β, ɾr, x, ɛ, ɪ, k^h, ɔ, œ, p^h, t^h, y, v

3. Experiments

To compare the models, we have performed three pair tests. For each test, 8 Japanese native speakers compared the quality and similarity to the original voice of 12 samples each.

To evaluate the similarity, subjects had to decide which of the synthesized voices was closer to the original one.

To evaluate the quality, we asked the subjects to focus mainly on the understandability, so that samples with better audio quality but more difficult to understand were rated lower.

多言語 HMM 音声合成における言語の追加
ハビエル・ラトレ、岩野 公司、古井 貞熙 （東工大）

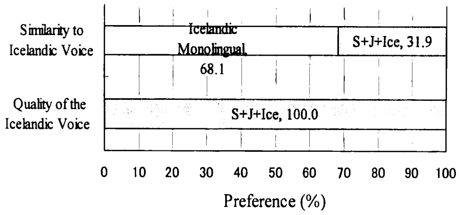

Figure 1: Preference score for a monolingual Icelandic vs trilingual model when adapted to an Icelandic voice.

3.1 Adaptation to an Icelandic speaker

In the first test, we wanted to verify that the results obtained for Spanish and Japanese were also valid for Icelandic, i.e. the multilingual model had better quality than the monolingual one for cross-lingual synthesis. For this, we have adapted the monolingual Icelandic model and the trilingual model (S+J+Ice) to the same Icelandic voice. With the adapted models, we synthesized Japanese texts. Figure 1 shows the results. Although the similarity to the original voice is still significantly better for the monolingual model, the quality of the trilingual model is overwhelmingly better.

3.2 Adaptation to Spanish and Japanese speakers

With the second test, we wanted to confirm that the inclusion of a third language did not deteriorate the quality of the synthetic speech. For this, we adapted the bilingual (S+J) and trilingual (S+J+Ice) models to the same Spanish and Japanese voices. For each voice, we compared the Japanese texts synthesized with both adapted models. The results can be seen in figure 2. We have not found any statistically significant difference between the two compared models neither in the speech quality nor in the the similarity to the original speaker.

3.3 Adaptation to a new language speaker

Finally, we compared the trilingual and bilingual models when adapted to speakers of languages not included in the training data, in our case English. With the inclusion of Icelandic, the total number of phones increased for 20%. Due to this increment, the English phonemes that have to be mapped to a phone with different IPA representation decrease from 10 to 6 within a total of 37 English phonemes, i.e. the phonetic coverage of English increases an 11.1 %. Figure 3 shows the results. The quality of the voice generated with the trilingual model was found to be 10% better, with a confidence ratio of 2.5%. The similarity to the target voice however, is not as good as we expected and presents no apparent difference between the two models.

4. Conclusions

Our proposed method of combining several monolingual corpora to create a polyglot HMM-based synthesizer has been applied to three languages that are not phonetically close without suffering a degradation of the quality.

Furthermore, the polyglot system adapted to a speaker in the new language can synthesize the other languages of the system with much better quality than a monolingual synthesizer in the new language that uses phone-mapping.

The inclusion of a new language improves the quality of the synthesized speech when the model is adapted to a language not included in the training data. This quality improvement is not followed by an improvement of the similarity to the target speaker that is still very poor.

5. Future Work

Most subjects commented that none of the presented samples really resembled to the original speaker. To improve this we want to apply speaker normalization techniques.

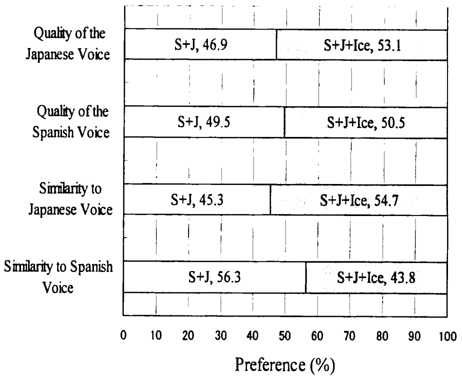

Figure 2: Preference scores for a bilingual vs a trilingual model when adapted to an Spanish and a Japanese voice.

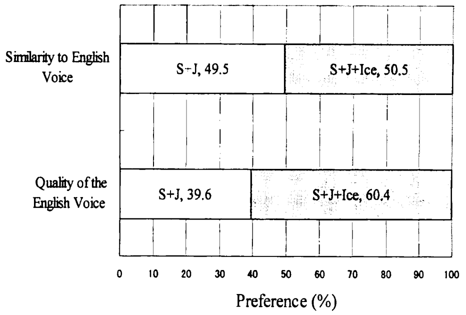

Figure 3: Preference scores of bilingual vs a trilingual model when adapted to an English voice.

In addition to adding new languages, we want to try some methods to synthesize languages for which we have very limited or no speech data. We want to try more accurate methods for phone-mapping and also test the feasibility of model interpolation instead of mapping.

6. Acknowledge

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.

References

[1] D. Graddol, "The Future of Language," *Science*, vol. 303, pp. 1329-1331, 2004

[2] T. Schultz et al., "Language Independent and Language Adaptive Acoustic Modeling," *Speech Communication*, Vol 35, Issue 1-2, pp 31-51, 2001

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

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

[5] J. Latorre, et al., "Toward an HMM-based Polyglot Synthesizer," *Proc. ASJ Fall meeting*, pp. 321-322, 2004

[6] T.Masuko, et al., "Speech Synthesis using HMMs with Dynamic Features," *Proc. ICASSP*, pp. 389-392, 1996

[7] M.Tamura, et al. "Speaker Adaptation for HMM-based Speech Synthesis System using MLLR," *Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pp. 273-276, 1998

[8] T.Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlruhe University," *Proc. ICSLP*, pp. 345-348, 2002