論文 / 著書情報
Article / Book Information

Title(English)	Progress on a Speaker Adaptable Polyglot Synthesizer
Authors(English)	Javier Latorre, Koji Iwano, Sadaoki Furui
Citation(English)	International Symposium on Large-Scale Knowledge Resources (LKR 2006), Vol. , No. , pp. 125-128
発行日 / Pub. date	2006, 3

Proc. International Symposium on Large-scale Knowledge Resources (LKR2006), 2006.3

Progress on a Speaker Adaptable Polyglot Synthesizer

Javier Latorre, Koji Iwano, Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology, Tokyo, Japan
{latorre,iwano,furui}@furui.cs.titech.ac.jp

Abstract

In this paper we present our method to synthesize multiple languages with any arbitrary voice. We call our approach, HMM-based speaker adaptable polyglot synthesis. The idea consists of using speech data from several speakers in multiple languages, to train a speaker and language independent acoustic Hidden Markov Model that can be adapted to imitate the voice of any given speaker. The model adaptation is done by means of Maximum Likelihood Linear Regression using some minutes of speech data from the selected target speaker. Using the adapted model it is possible to synthesize any of the languages used to train the speaker independent model with the voice of the selected speaker, regardless of the language actually spoken by that speaker. The results of a subjective evaluation show that when the language of the original speaker and the language to be synthesized are different, the performance of our method is better than that of other methods based on monolingual acoustic models and phone mapping. Also in the case of languages not included in the training data, the performance of our approach equals or surpasses that of any of the monolingual synthesizers built in the languages used to train the multilingual acoustic model.

1. Introduction

As a side effect of the globalization, there has been a dramatically increment in the number of people that have to use two or more languages in their daily life. If we consider this, it is normal to expect that these people will soon demand software applications that can deal with multiple languages in the same way they do. Moreover they will demand applications that help them with those languages they cannot speak fluently.

The goal of this research is the development of a system able of synthesizing multiple languages with the same voice and where we can modify the output voice to imitate the voice of any arbitrary speaker.

Although it is possible to synthesize each language with a different voice, in applications that have to deal with mix-lingual text such as e-mail reading, this approach does not seem appropriate.

At the same time, many speech synthesis applications will require the ability of personalizing the output voice, that is, to transform the output voice into the voice of the user or any other new speaker without recording much speech data from that speaker. From a multilingual perspective, this implies in many cases that the language spoken by the original speaker and the language we want to synthesize will be different, for example in a speech-to-speech translator.

Besides the correct synthesis of mix-lingual text, another application of our system is the fast and economical implementation of speech synthesizers for minority languages. To reduce cost in the development of a speech synthesizer for a new languages with very limited or no available speech data, the most normal approach is to use speech data from a similar language and apply a set of phone mapping rules. The main problem is that this mapping introduces an error which varies with the phonetic similarity between the target and the substitute language. Our assumption is that by combining several languages, since the palette of available sounds is wider than when only one language is used, it is possible to find more appropriate substitutes for the sounds of the target language. In this way, we can reduce the mapping error and thus improve the quality of the synthesized speech.

2. Previous approach to polyglot synthesis

The approaches to polyglot synthesis existing so far consist in a) recording speech data from a human polyglot speaker, or b) mapping the phones of the target language onto the phones of the language for which the synthesizer was built.

The first approach [1], can provide the quality of state-of-the-art unit selection synthesizers. However, to find good voice talents for more than 3 languages is always a difficult challenge. Moreover the system is by definition limited to the languages spoken by that polyglot speaker.

The second approach [2], can be applied to any language or speaker. However, the resulting voice retains a very strong foreign accent. As a result, when the target and substitute languages are phonetically very different the synthetic voice becomes almost unintelligible.

3. HMM-based speech synthesis

In [3], we proposed a method to create a polyglot synthesizer, based on HMM synthesis [4]. It consists in mixing speech data from several monolingual speakers for each one of the languages we want to synthesize. Our hypothesis is that voice differences depends only on anatomical factors. Consequently, the average voice created by mixing a sufficient number of speakers should be almost the same for any language. The architecture of our proposal is shown in Fig. 1. Our approach has three main steps: training, adaptation and synthesis.

In the first step, we train a speaker independent (SI) HMMs by mixing speech data from several speakers in multiple languages. To train these models, first we normalize the phonetic representation of all the languages we want to mix. In this way we obtain a reduced set of labels where each label refers to the same phone. Then, a set of triphone models with the same topology is trained. After that, we perform a state by state clustering of the triphone models using a single decision tree for all the phones so that we can take advantage of the similarities across phones [5]. The questions of this tree refer only to the phonetic features of the phones but not to the phones themselves.

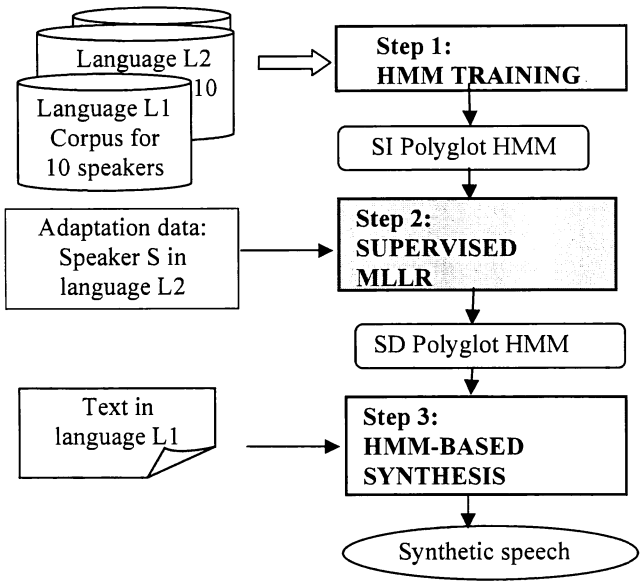

Figure 1: Architecture of an HMM-based polyglot synthesizer in the case of cross-lingual synthesis.

In the second step, the SI model trained in step one is transformed into a speaker dependent (SD) model with the voice of the target speaker. This is done by means of MLLR [6] using some minutes of speech data from that specific speaker. In our approach, the use of adaptation is mandatory not only to personalize the output voice but also to obtain a homogeneous voice across languages. Otherwise, since the phonetic coverage of the mixed speaker is not equal the voice identity could be different from language to language or even change abruptly from one phone to the next in the middle of a word.

The level of similarity between the original speaker and the speech synthesized with the SD model depends, among other factors, on the number of adaptation classes that is used. Generally speaking, the more classes are used the more similar to the original speaker. However, an excessive number of adaptation classes degrades the intelligibility of the synthesized speech. The optimum number of classes depends mainly on the number of final leaves of the SI model and the amount of available adaptation data.

In the last state, HMM-based synthesis, the adapted HMMs are concatenated according to the sequence of phones of the input text. Then for this sequence of HMMs, the sequence of output parameters with higher likelihood is calculated [6]. The final waveform is generated by convoluting the output parameters with a source excitation signal created according to the input text.

In our polyglot synthesizer, all the languages that were included in the training data can be synthesized directly. However for extrinsic languages which were not included in the training, first we need to apply mapping between the phones of the new language and those of the included languages. In this case, since the set of phones that we obtain by combining multiple languages is wider than the set of phones that can be obtained from a single language the mapping error is lower which results in better intelligibility.

4. Experiments

In order to compare our approach with other possible approaches based on phone mapping we have performed a subjective evaluation. The evaluation parameters are:

a) Perceptual Intelligibility: how easy it is for subjects to understand the synthesized speech.
b) Similarity between the synthesized voice and the original speaker
c) Native accent: whether the synthesized speech sounds like a native speaker or a foreigner.

Our goal was to test the performance of our polyglot synthesizer in three scenarios:

(1) *Cross-lingual synthesis*, the language to be synthesized and the language of the target speaker are both included in the training data but are different.
(2) *Synthesis of extrinsic languages*, the language of the target speaker is included in the training data but the language to be synthesized is not.
(3) *Direct synthesis*, the language to be synthesized and the language of the original speaker are the same and included in the training data.

Although the evaluation was performed only for Japanese and Spanish, the results can be extrapolated to other languages.

4.1. Acoustic models

In previous evaluations, we tested our method only for two phonetically close languages, Spanish and Japanese[3]. With this new evaluation we wanted to test whether our method can be also applied to phonetically distant languages. For this purpose, we trained several monolingual and multilingual acoustic models using speech data in Spanish, Japanese, German, French and Russian. Depending on the languages mixed in the model, they can be classified into three groups:

a) monolingual models
b) Polyglot models mixing Spanish and a two language combination of Russian, French, and German
c) Polyglot models mixing Japanese and a two language combination of Russian, French, and German

Each model was trained with approximately the same amount of data from 30 monolingual male speakers. In the polyglot models these 30 speakers were equally distributed among the three languages included in the model. For each language we selected from the database those speakers whose voices we found relatively similar one another.

All the models were tied-state triphone models with 1 Gaussian, 3 states left-to-right without skips. The feature vector consisted of the total energy, the 25 first mel-cepstral coefficients and their delta. The analysis window was a 16ms Blackman window with a 5 ms shift.

For speaker adaptation we used supervised MLLR. All the models were adapted to two speakers of each language used to train that particular model. Additionally, the Spanish and Japanese monolingual models were also adapted to two speakers of each one of the other languages. To be able to use supervised adaptation in this case, before applying the MLLR adaptation algorithm we mapped the phones of the labels of the adaptation data into the Spanish and Japanese phonetic sets respectively.

Due to the high number of models that we wanted to evaluate, it was almost impossible to test every possible combination of model size and number of adaptation classes,

so for each model we pre-selected the combinations of these two factors that yielded the better trade off between intelligibility and similarity to the original speaker. For speakers of languages included in the training data, we pre-selected models adapted with 4 classes. For speakers of languages not included, models adapted with 2 classes yielded the best result. The size of the SI models was decided using the MDL criterion []

Both training and adaptation data were selected from the Globalphone speech database[7]. Although this database was not specifically designed for speech synthesis, it was the best available database for fulfilling our need for multilingual data and multiple speakers for each language .

4.2. Phone mapping for the extrinsic language case

In order to synthesize speech in languages not included in the training data of the acoustic models, we used phone mapping. A mapping table was defined for each acoustic model. The rules to define these tables were:

a) Whenever possible, each Spanish and Japanese phone was substituted by phones represented by the same IPA symbol.
b) Otherwise, Spanish and Japanese phones were substituted by the available phone with the highest phonetic similarity, calculated from the articulatory feature vector.
c) If there exist more than one phone with the same similarity, the selected phone was the one that in the original language (Spanish or Japanese) is an allophonic variant of the phone we want to map.

To focus only on the acoustic models, we used original prosody extracted from the recorded audio version of the test texts. We approximated the prosody to the original speakers, by shifting the mean F0 of each test file to the mean F0 of each original speaker.

4.3. Experimental conditions

For the evaluation, 18 Spanish and 18 Japanese texts were synthesized by each one of the adapted models. These files were presented to 6 native Spanish and 6 Japanese native subjects respectively. The stimuli were divided among the subjects so that each subject listened to three stimuli for each adapted acoustic model.

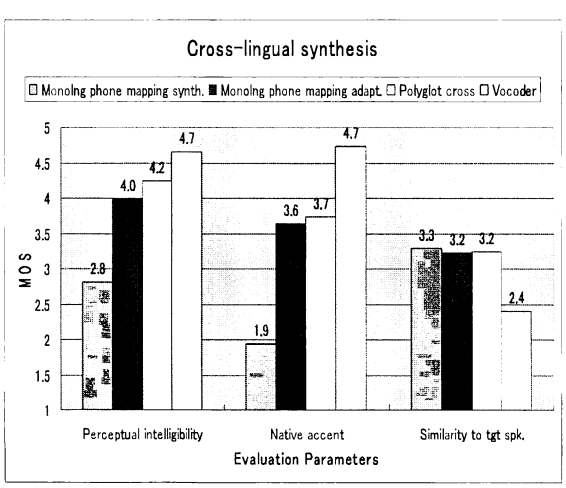

Figure 2: Performance for cross-lingual synthesis.

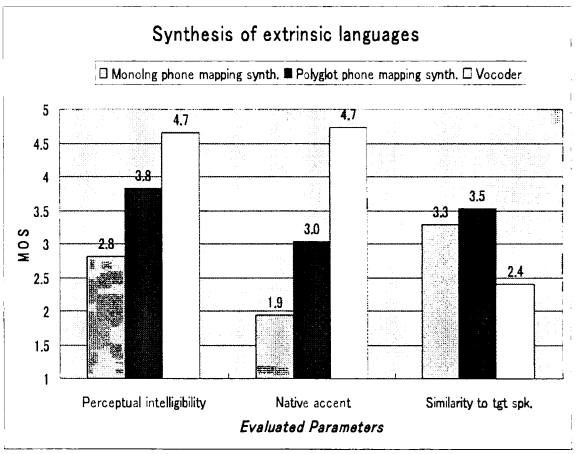

Figure 3: Performance for the synthesis of extrinsic.

Due to the high number of stimuli that each subject had to evaluate we distributed the experiment in 3 sessions. The distribution was made in such a way that at each session at least one file from each model were listened but the same test text were not repeated more than 4 times. In order to obtain uniform scores across sessions we also included at each session the vocoder resynthesis of the audio version of the test texts. The set of stimuli corresponding to one session were presented in a random order.

All three evaluation parameters were evaluated simultaneously on a 5 point MOS scale. To evaluate the similarity to the original speaker, subjects were asked to compare the synthesized speech with a short audio file of around 8 seconds with the voice of the original speaker.

5. Results

Figures 2, 3 and 4 show the results of the evaluation. The columns named "Monolng. phone mapping synth." represent the average scores of monolingual models using phone mapping to synthesize the target language. The columns named "Monolng. phone mapping adapt." represent the average of the Spanish and Japanese models adapted to speakers of other languages by means of phone mapping and used to synthesize Spanish and Japanese respectively. The columns named "Monolng. direct" represent the average score of monolingual models in direct synthesis. The columns named "Polyglot cross", "Polyglot phone mapping synth", and "Polyglot direct" represent the average score of the polyglot models in cross-lingual synthesis, synthesis of extrinsic languages, and direct synthesis respectively. The columns named "Vocoder" represent the scores obtained by the vocoder resynthesis. This column is the ceiling for perceptual intelligibility and native accent, and the noise level for similarity.

Figure 2 shows the performance of the different models in the case of cross-lingual synthesis. It can be seen that the perceptual intelligibility and native accent obtained with the polyglot models is clearly superior to that obtained with the monolingual models and phone mapping for synthesizes. The difference between the polyglot models and the monolingual model with cross-lingual adaptation is not so great, although in terms of perceptual intelligibility this difference is still significant. With respect to the perceived similarity to the

original speaker the scores of the three models were basically the same.

Figure 3 shows the result in the case of synthesis of extrinsic languages. In this case both models use phone-mapping for synthesis. However, as we can see the performance of the polyglot models is still clearly superior to that of the monolingual ones. Indeed, the performance of the polyglot models in this case is always equal to or better than the performance of the best monolingual model in any of the languages used to train the polyglot model. This result confirms our idea that the mapping accuracy and with it the perceptual intelligibility can be improved by using polyglot models instead of monolingual ones.

Figure 4 shows the results for direct synthesis. Due to the fact that in the polyglot model we are mixing different languages, we expected that in the case of direct synthesis its performance would be worse than that of a monolingual model. However, the scores for the polyglot model were just slightly worse than that of the monolingual models, and in terms of perceptual intelligibility this difference was not statistically significant.

In Fig. 4 we also include the scores of the "Polyglot cross" models. We can see that although the performance of the polyglot models in cross-lingual synthesis is clearly worse than in direct synthesis, the differences are small especially if compared with the differences between the polyglot model and the monolingual ones in cross-lingual synthesis or in the synthesis of extrinsic languages.

6. Conclusions

In this paper we have presented our approach to polyglot synthesis and the results of a subjective evaluation performed for Japanese and Spanish. We have shown that it is possible to build a polyglot synthesizer by mixing the data of monolingual speakers in different languages even in the case of languages which are not phonetically similar. Furthermore, since our approach is based on HMM-synthesis, we can easily adapt the system to imitate the voice of any given speaker.

For cross-lingual synthesis we have shown that our system performs much better than other approaches based on a phone mapping. The performance of the approach based on cross-lingual adaptation of a monolingual model was not so different much from that of the polyglot models. However, to synthesize multiple languages with this method we need to adapt as many monolingual models as languages we want to synthesize. Since each language is trained separately, the risk of having different output voices for each language is higher than when all the languages are trained together into a single polyglot model. With our approach, we can synthesize without voice change and with very similar quality all the languages included in the training data. In the case of direct synthesis, this quality does not differ significantly from that obtainable with a monolingual acoustic model.

Additionally, for the synthesis of extrinsic languages, the performance levels that we can obtain using a polyglot model are better than those obtained with monolingual models trained with any of the languages used to train the polyglot model. This result makes our approach especially attractive for minority languages, for which the amount of available speech data is usually either very limited or inexistent.

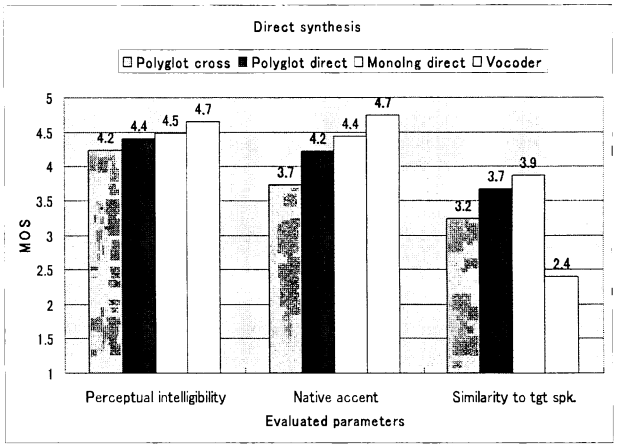

Figure 4: Performance for direct synthesis.

7. Future work

Our next goal is to test different phone mapping methods for the synthesis of extrinsic languages. Additionally we want to try if it is possible to use some kind of phone interpolation instead of phone mapping. By interpolating between different triphone models we expect to reduce the mapping error and thereby improve the speech quality.

We also want to explore which solution may be used to predict prosody, as well as a possible means of interlacing the prosody of multiple languages, as might be necessary for texts which contain words from more than one language.

8. References

[1] Traber, C., Huber, K., Nedir, K., Pfister, B., Keller, E. and Zellner, B., "From multilingual to polyglot speech synthesis", *Proc Eurospeech*, pp.835-838, Budapest, Hungary 1999.

[2] Campbell, N., "Talking foreign. concatenative speech synthesis and the language barrier", *Proc. Eurospeech*, pp. 337-340, Aalborg, Denmark 2001.

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

[4] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., Imai, S., "An algorithm for speech parameter generation from continous HMMs with dynamic features", *Proc. Eurospeech*, pp. 757-760, Madrid, Spain 1995.

[5] Yu, H., Schultz, T., "Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition", *Proc. Eurospeech*, pp. 1869-1872, Geneve, Switzerland 2003.

[6] Schultz, T., "Globalphone: a multilingual speech and text database developed at Karlsruhe university", *Proc. ICSLP*, pp. 345-348, Denver, USA 2002.

[7] Tamura, M., et al., "Text-to-speech synthesis with arbitrary speaker's voice from average voice", *Proc. Eurospeech*, pp. 345-348, Aalborg, Denmark 2001.

[8] Shinoda, K., Watanabe, T., "MDL-based context-dependent subword modeling for speech recognition.", *Journal of the Acoustic Society of Japan* (English), vol. 21, pp. 79-86, Mar. 2000