論文 / 著書情報
Article / Book Information

Title(English)	Speaker adaptable multilingual synthesis
Authors(English)	Javier Latorre, Koji Iwano, Sadaoki Furui
Citation(English)	Symposium on Large-Scale Knowledge Resources (LKR2005), Vol. , No. , pp. 235-238
発行日 / Pub. date	2005, 3

SPEAKER ADAPTABLE MULTILINGUAL SYNTHESIS

Javier Latorre, Koji Iwano, Sadaoki Furui

Tokyo Institute of Technology,
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{latorre,iwano,furui}@furui.cs.titech.ac.jp

ABSTRACT

In this paper, we propose a new approach to polyglot synthesis. It is based on an HMM synthesis technique. We combine several monolingual corpora from different languages to create an average voice. Using this voice, it is possible to synthesize speech in any of the languages of the training corpora with the same voice individuality. Furthermore, the average voice can be modified to imitate the voice of any real speaker. So we can synthesize several languages with the voice of any arbitrary speaker, regardless of the language spoken by that speaker.

1. INTRODUCTION

Since the end of WWII, English has become the main language for international communication. However, the economical and demographic changes are modifying this scenario. According to some previsions [1], in 2050 the number of Spanish, Arabic and Hindi speakers will have reached the number of English speakers. No doubt that English will still be the main common language, but even in that case, most people will probably need to speak two or more languages in their daily life. In such context, it is logical to think that there will be a strong demand for technology that helps to interact with other people in foreign languages without needing to study those languages.

One of the main assets of speech synthesizers is their capacity to generate many different languages. To achieve this, current speech synthesizers usually switch the output voice. However, for some applications this solution might not be appropriated because it implies a change of the individuality of the output voice. Such applications have to be not only multilingual but polyglot.

2. BACKGROUND

Our purpose is to create a system that can synthesize many languages with any arbitrary voice. In other words, the system should be able to, synthesize Spanish and Japanese, for example, with the voice identity of an English speaker who does not speak neither of those languages. To achieve our goal we need polyglot synthesis and speaker adaptation.

2.1 Polyglot synthesis

There were two traditional approaches toward polyglot synthesis. Thy both are based on concatenative synthesis.

The first one [2], consist of recording speech data in several languages from a human polyglot speaker. In this way, it is possible to generate speech with the quality of unit selection synthesis. However, to find a good polyglot voice talent can be extremely difficult. Moreover, this method can only be applied to a limited number of languages.

The second approach [3] consists of mapping any foreign sound, to the closest native sound. This method works pretty well when the foreign and the native languages are phonetically close. However, the resulting voice retains always a strong foreign accent. If the two languages are not phonetically close, this accent degrades the understandability. Another problem is that the resulting speech in the mapped language is more chopped than in the native one due to an increment of the number of concatenation points. This occurs because the combinations of phonemes in the target language are usually uncommon in the original one.

2.2 Speaker adaptation

The most usual case for our system is that the user cannot speak the language he wants to synthesize, so we need cross-lingual voice conversion. Mashimo et al. [4] showed that cross-lingual voice conversion using GMM is possible. They obtained almost the same performance for voice conversion across languages than for conversion within the same language. However, in their approach at least one bilingual database to train the voice mapping was required.

3. HMM-BASED POLYGLOT SYNTHESIS

Our system is based on an HMM synthesis technique [5]. The voice quality of HMM synthesis is lower than the quality of unit selection synthesis. However, it provides the flexibility in voice conversion and prosody modification that we need. The HMM synthesis has three phases. First, a set of HMMs is trained with the speech database of one or more speakers. Then, the models are adapted to a given speaker. Finally, the text to be synthesized is transformed into a sequence of models from which the speech parameters are calculated. Figure 1 shows the whole process for Spanish and Japanese monolingual HMMs and for a bilingual HMM when they are adapted to a Spanish speaker's voice and synthesized a text in Japanese.

3.1 Database

As training data we have used the Globalphone corpus [6] for Spanish and Japanese and the Jenssons's corpus for Icelandic. We would like to remark, that the mentioned corpora are not specifically designed for speech synthesis.

We have trained the models with 10 minutes of data from ten speakers for each language. The training data altogether is approximately 6 hours. As adaptation data for the integrated languages, we have used 10 minutes of speech from speakers not included in the training. For the adaptation to an English speaker, we have used 10 minutes of speech from the "awb" voice of the Arctic corpus [7].

3.2 Models training

To train the polyglot HMMs, first we converted the transcriptions of the training data to a common multilingual format. We assigned the same label to sounds that share the same IPA code.

Second, we trained triphone models and clustered them with a phonetic decision tree. We used the same decision tree for all the triphones so that the parameters of different phones can be mixed [8]. The questions of the decision tree were exclusively about the articulatory features of the triphones. The introduction of questions about the language to which the phones belong did not produce any noticeable improvement.

3.3 Speaker adaptation

The problem of combining several speakers' voices into a speaker independent model is the lack of individuality of the resulting voice. Moreover, if the phone coverage is not equal for all the speakers, this individuality might change abruptly in the middle of an utterance. In our system it is impossible to have the same phone coverage for all the speakers because we are mixing speakers and languages. To improve the coherence of the output voice, we transformed the speaker independent voice into the voice of a specific speaker by means of supervised MLLR adaptation [9]. Since the adaptation of the variances produced unnatural values we only adapted the mean values.

Generally speaking, the similarity to the original speaker increases with the number of adaptation matrices. However, too many matrices degrade the quality of the synthetic speech especially when the language to be synthesized and the language of the target speaker are not the same. To find the optimum number of adaptation matrices, we adapted the HMMs with different number of matrices and pre-selected the adapted models with the best trade-off between similarity to the target speaker and speech quality.

For speakers of languages included in the training data the adaptation of the HMMs is done directly. For speakers of not included languages we used phone-mapping. This was the case of cross-lingual adaptation of the monolingual models and the adaptation of the polyglot model to an English speaker. This mapping was done by rules. Basically we mapped each external phone onto the phonetically closest internal one.

3.4 Synthesis

To synthesize a text as speech, first its phonetic transcription is converted into a sequence of HMMs. If the phonemes to be synthesized are included in the training data, this conversion can be done directly.

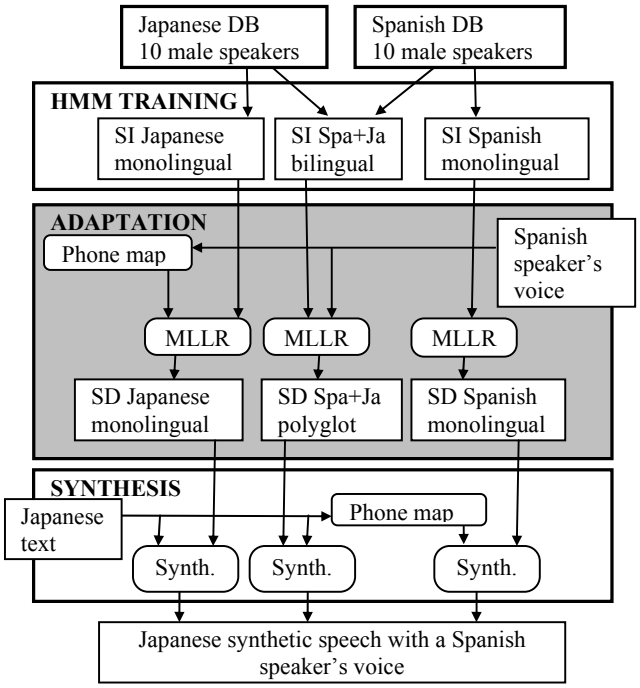

Fig. 1 HMM training, adaptation and synthesis

Otherwise, first the foreign phones have to be mapped onto the correspondent native ones. For synthesis we used the same set of mapping rules as for cross-lingual adaptation.

4. EVALUATION

To evaluate the performance of our system in terms of subjective intelligibility and similarity to the target speakers, we have performed two set of experiments. In both cases the subjects were Japanese native speakers, and the texts synthesized were in Japanese.

To focus only on the cepstral information, we used the original prosody of the evaluation files. To adapt the prosody to the target voices, we modified the mean pitch and pitch dynamic of the evaluation files to the characteristics of the target speakers.

4.1 Comparison of a bilingual versus monolingual models

In the first set of experiments, we have evaluated the performance of a bilingual model (Spanish and Japanese) against a Spanish and a Japanese monolingual models. To get a referenced, we have also evaluated a Spanish diphone concatenation synthesizer and a vocoder reconstruction of the original utterance.

In these experiments we asked the subjects to score the intelligibility and similarity to the target speaker of the evaluation utterances in a 5 points scale (1-very poor, 5-very good)

For these experiments we have used 3-states left-to-right triphone models without skips. The feature vector consists of 25 mel-cepstral coefficients and their delta coefficients. Each state was modeled by 4 Gaussians. The transitions between states were modeled by state transition matrices. The data were windowed by a 30 ms Blackman window with a 5 ms shift.

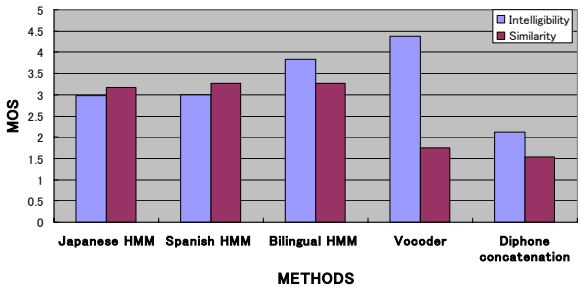

Fig. 2 Results for models adapted to Spanish speakers

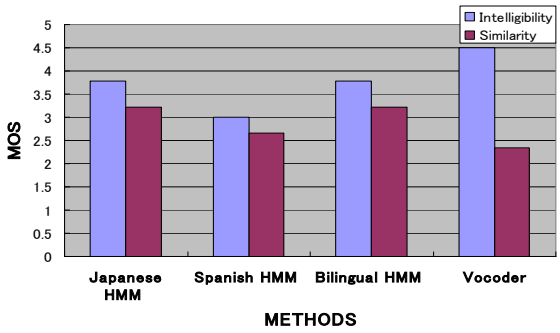

Fig. 3 Results for models adapted to Japanese speakers

4.2 Addition of a new language to the system

In the second set of experiments we analyzed the effect of adding a phonetically distant language (Icelandic) to the Spanish-Japanese bilingual model. We did three pair tests to compare the performance of:

a) the trilingual HMMs versus the Icelandic monolingual HMMs when adapted to an Icelandic voice

b) the trilingual versus the bilingual HMMs when adapted to Spanish and Japanese voices

c) the trilingual versus the bilingual HMMs when adapted to a speaker of an external language (English)

In these experiments we have used the same HMM configuration as in the previous experiments but with some modifications. We have decreased the length of the analysis window from 32ms to 16 ms to avoid that the minimal time required by the sequence of HMMs exceeds the real time of the sound, especially in the case of combined sounds such as diphthongs. We have also reduced the number of mixture from 4 to 1 to limit the processing time.

5. RESULTS

5.1 Combination of phonetically close languages

Figure 2 shows the results of the subjective intelligibility test for models adapted to Spanish voices The subjective intelligibility of the polyglot model outperforms the monolingual models and lies in between those and the vocoder reconstruction which is the maximal value that can be obtained. In this experiment, we found no significant difference in the similarity to the target speaker of the bilingual and the monolingual HMMs

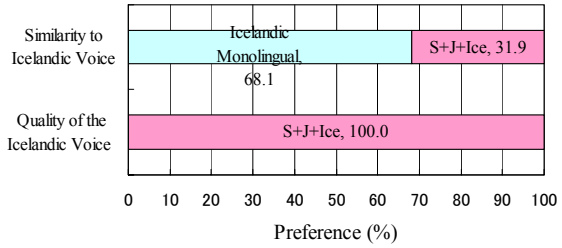

Fig. 4 Preference score for a monolingual Icelandic vs trilingual model when adapted to an Icelandic voice

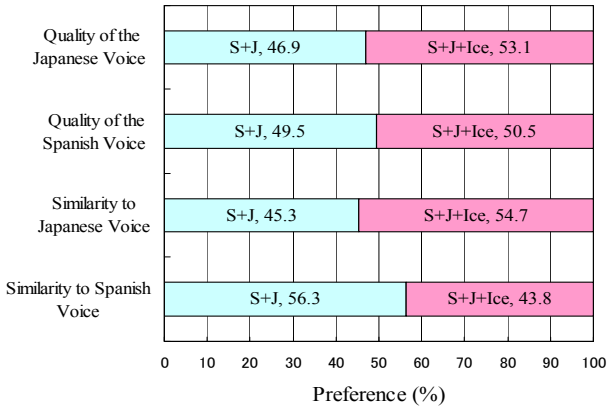

Fig. 5 Preference scores for a bilingual vs. a trilingual model when adapted to an Spanish and a Japanese voice.

Figure 3 shows the results of the HMMs adapted to Japanese voices. We found no significant difference between the subjective intelligibility and similarity of the bilingual and the Japanese monolingual models. This means, that the combination of two languages does not produce any degradation of the intelligibility.

There is a significant difference of the performance between the Spanish monolingual model when adapted to Japanese voices to Spanish voices. This is due to the evaluation method and the double phone-mapping at adaptation and synthesis. When subjects compare voices in two different languages, they can only concentrate on the characteristics of the voice. Other features of the voice individuality such as prosody or accent are a priori assumed to be different. However, when we presented to the subjects the utterance synthesized with the Spanish monolingual model adapted to Japanese voices, they were listening to Japanese spoken by a native Japanese voice and Japanese spoken by an Spanish accented voice. In such cases, unless the similarity is very clear subjects always tend to consider these voices as different

The subjective intelligibility of the Japanese model is significantly better when adapted to Japanese voices than to Spanish voices but there is not such difference for the Spanish model in spite of the double phone-mapping.

5.2 Combination of phonetically distant languages

Figure 4 shows the results of the comparison between the trilingual model (Spanish+Japanese+Icelandic) and the Icelandic monolingual model when adapted to an Icelandic speaker. With this 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. Although the similarity to the original voice of the monolingual model is still significantly better, the quality of the trilingual model was preferred by all the subjects.

Figure 5 shows the result of the comparison between the bilingual (Spanish+Japanese) and the trilingual models when adapted to Spanish and Japanese speakers. With this test we wanted to confirm that adding a phonetically distant language does not deteriorate the quality of the synthetic speech. We have not found any significant difference between the two compared models neither in the speech quality nor in the similarity to the original speaker. Apparently, the integration of a third language produces no degradation of the performance.

Figure 6 shows the results of comparing the trilingual and bilingual models when adapted to speakers of languages not included in the training data, in this case English. It can be seen that the quality of the voice generated with the trilingual model was found to be 10% better (confidence ratio of 2.5%). The inclusion of Icelandic produces an increment in the number of phones of the system of a 20%. Thus, the English phonemes that need to be mapped to a different phone, i.e. a phone with different IPA representation, decrease from 10 to 6 within a total of 37 English phonemes, i.e. the phonetic coverage increases. In spite of this, there is no apparent difference in the similarity to the target voice.

5. CONCLUSIONS

We have proposed a new method for polyglot synthesis that combines monolingual corpora to create an average polyglot voice. With this voice it is possible to synthesize speech in any of the languages of the training corpora without changing the voice individuality. Moreover, the individuality of this average voice can be adapted to imitate any arbitrary speaker.

Our method works equally well for two and three languages. Indeed, the inclusion of a new language seems to improve the quality of the speech when the model is adapted to speakers of languages not included in the training data.

In the case of adaptation to Spanish or Icelandic speakers and synthesis of Japanese texts, the subjective intelligibility of the proposed method outperforms any monolingual methods based on phone-mapping. For Japanese target speakers the subjective intelligibility of the polyglot system is the same as a Japanese monolingual one. Informal tests suggest similar results could be achieved for Spanish and Icelandic.

The MOS scores of the similarity between the synthetic voices and the target voices of the monolingual methods and the proposed polyglot one are not significantly different. However, when both models are compared in a pair test, the monolingual one was preferred

Since the size of the polyglot model is always smaller than the combined size of the monolingual models our approach can be useful in multilingual applications that require minimal footprint.

The evaluation shows that for cross-language synthesis HMM-based methods outperform diphone concatenation with phoneme mapping.

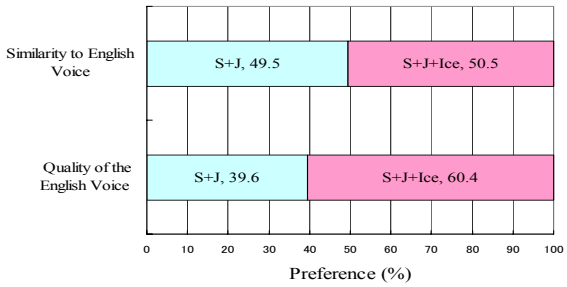

Fig. 6 Preference scores of bilingual vs. a trilingual model when adapted to an English voice.

6. FUTURE WORK

Our main goal is to improve the quality of the synthetic speech and the similarity to the original voice while increasing the number of languages that can be synthesized. To achieve this, we will test different clustering methods and central voice techniques.

In addition to adding new languages, we want to experiment with methods to synthesize languages for which there are very limited or no speech data. This would be the case of minority languages. We want to try more accurate methods for phone-mapping based on the decision tree. We also want to test the feasibility of model interpolation instead of mapping for not existing sounds.

Finally, we plan to record a bilingual or pseudo-bilingual corpus. Including such a corpus in the training data should increase the cohesion of the polyglot voice. This corpus should also be useful to evaluate cross-lingual adaptation.

REFERENCES

[1] D. Graddol, "The future of language," *Science*, vol. 303, pp. 1329-1331, Feb. 2004

[2] C. Traber et al., "From multilingual to polyglot speech synthesis," *Eurospeec99*, pp.835-838, Sept.1999

[3] N. Campbell, "Talking foreign. Concatenative speech synthesis and the language barrier," *Eurospeech01*, pp 337-340, Sept. 2001

[4] M. Mashimo et al., "Evaluation of cross-language voice conversion based on GMM and STRAIGHT," *Eurospeech01,* pp. 361-364, Sept. 2001

[5] T.Masuko et al., "Speech synthesis using HMMs with dynamic features," *ICASSP-96*, pp. 389-392, May 1996

[6] T.Schultz et al., "The GlobalPhone project: multilingual LVCSR with Janus-3," *Multilingual Information Retrieval Dialogs:2nd SQEL Workshop*, Apr. 1997

[7] J. Kominek et al., "The CMU Arctic Speech Databases," *5th ISCA Speech Synthesis Workshop*, pp. 223-224, Jun. 2004

[8] H. Yu et al., "Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition,", *Eurospeech03*, pp. 1869-1872, Sept. 2003

[9] M.Tamura et al., "Speaker adaptation for HMM-based speech synthesis system using MLLR," *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 273-276, Nov. 1998