POLYGLOT SYNTHESIS USING A MIXTURE OF MONOLINGUAL CORPORA

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

This paper proposes a new approach to multilingual synthesis based on an HMM synthesis technique. The idea consists of combining data from different monolingual speakers in different languages to create a single polyglot average voice. This average voice is then transformed into any real speaker's voice of one of these languages. The speech synthesized in this way has the same intelligibility and retain the same individuality for all the languages mixed to create the average voice, regardless of the target speaker's own language.

1. INTRODUCTION

English is the indisputable international language for business and communication. However, other languages such as Spanish, Japanese or Chinese are becoming more and more important [1]. Due to the globalization, the number of people that have to use two or more languages in their daily life is growing and with them the number of applications that require multilingual capacity. Two examples of this fact are the 30 million Spanish speakers already living in the USA and the 25 official languages in the EU.

The most attractive possibility of speech synthesizers is to speak many languages. Although multilingual synthesis can be achieved by switching the language module, this solution might not be appropriate in most cases, especially if it implies a change in the individuality of the output voice.

Another application of multilingual technology is preserving endangered languages. Multilingual systems can reduce the implementation costs of speech technology by reusing the resources collected for other languages [2].

2. BACKGROUND

The goal of our research is to create a multilingual synthesizer than can speak any language with any given voice. Such synthesizer should be able to, for example, synthesize with the same quality Spanish and Japanese text with the voice of a Japanese speaker who does not speak Spanish. To achieve this goal we need on the one hand the capability of multilingual synthesis and on the other the voice conversion.

Traber et al. [3] proposed a distinction between polyglot and multilingual systems. They defined "polyglot systems" as those that can synthesize several languages using the same voice with appropriate pronunciation, and "multilingual systems" as those that have to change the synthesis process and output voice to synthesize different languages.

In this paper we will call by polyglot a system that can generate intelligible speech in several languages having the same voice identity.

2.1 Multilingual synthesis

The traditional approaches to polyglot synthesis are based whether on recording a multilingual corpus from a polyglot speaker [3] or on mapping the phones of the foreign language onto the phones of the original one [4].

With concatenative synthesis, the first method can produce high quality speech. However, finding good voice talents for more than 3 or 4 languages or for rare combinations of 2 languages is very difficult. Furthermore, such systems are hardly expandable.

The second method has been successfully applied to phonetically close languages [5]. The resulting speech is understandable but retains the foreign accent of the original speaker. Foreign accent is not necessarily bad and in some cases can even improve the acceptability of the synthetic voice [6]. However, if the accent gets too strong, the intelligibility decreases significantly. Another problem of phone-mapping in concatenative synthesis is that the resulting sequence of phones tends to be uncommon in the original language. This makes the synthetic speech chopped and discontinuous.

2.2 Voice conversion and speaker adaptation

Some applications require transformation of the output voice i.e. its individuality, without recording much new data. In speech recognition, voice adaptation techniques based on the MLLR or the MAP algorithms are well known. These techniques have also been applied to the HMM speech synthesis [7].

In a multilingual scenario, there are some applications that would benefit from voice conversion, for example a speech-to-speech translator. For this example the most usual case is that the user cannot speak the language he wants to synthesize. This implies that cross-lingual voice conversion is needed. Mashimo et al. [8] showed that cross-lingual voice conversion using GMM is possible. According to their results, the performance for voice conversion across languages was nearly the same as for conversion within the same language. However, their approach requires at least one bilingual database to train the voice mapping.

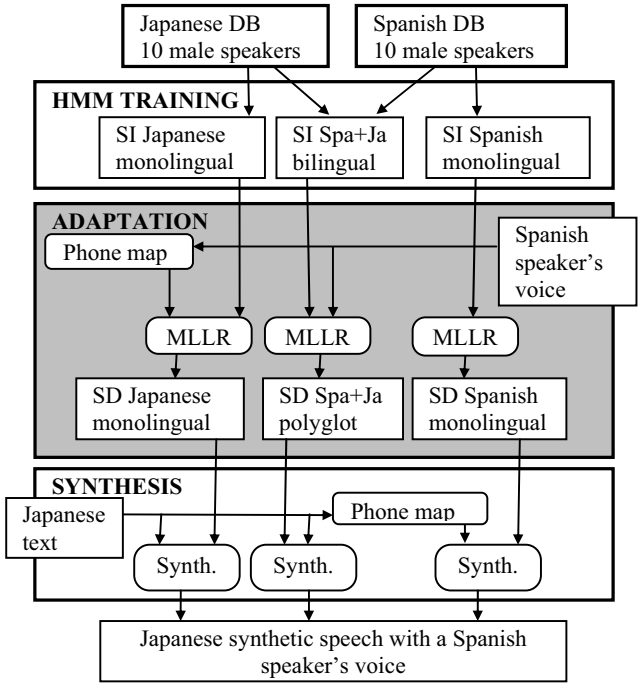

Fig. 1 HMM training, adaptation and synthesis

2.3 Multilingual phonetic transcription

A multilingual synthesizer needs a coherent and language independent phonetic representation. IPA assigns a unique symbol to each combination of articulatory features i.e. to each phone. There are several computer-readable versions of IPA. SAMPA, the most popular one, is not good for multilingual applications because it usually assigns the same ASCII code to different IPA symbols for different languages. Worldbet [9] is more consistent. It codes each IPA symbol differently and provides an easier way of coding complex sounds.

3. OUR APPROACH

Our system is based on an HMM synthesis technique [10]. Although HMM synthesis cannot produce the quality of unit selection synthesis, it provides the flexibility in voice conversion and prosody modification that we need. The HMM synthesis has three phases. In the first one, a set of HMMs is trained with the speech database of one or more speakers. In the second, the models are adapted to a given speaker by means of speaker adaptation techniques. In the last phase, the text to be synthesized is transformed into a sequence of adapted models from which the speech parameters are generated.

Although our final goal is to synthesize any language, in a first step we have limited our research to two phonetically close languages: Spanish and Japanese. We have selected these two languages because they share around 60% of their phones, there are available monolingual speech resources for both of them and we know them both reasonably well.

In our method, first we trained speaker independent polyglot HMMs, mixing data from Spanish and Japanese speakers. Then we adapted these SI models to several Spanish and Japanese speakers. Finally, with the adapted voices, we synthesized Japanese and Spanish texts.

To compare the performance of our approach, we trained two monolingual speaker-independent HMMs for Spanish and Japanese. These monolingual SI models were adapted to the same Japanese and Spanish voices as the polyglot model and used to synthesize the same Spanish and Japanese texts.

Figure 1 shows the whole process for the three described models when they are adapted to a Spanish speaker's voice and synthesized a text in Japanese.

3.1 HMM training

To create the HMMs, the transcriptions of the training data were first converted to a modified version of Worldbet. To allow a distinction between Japanese and Spanish phones in the early stages, we added a language tag to all the symbols except silence ones.

Second, we clustered the states of the models with a phonetic decision tree. We used a single clustering tree for all the phones so that parameters could be shared across phones [11]. The questions to construct the clustering tree were about the phonetic features of the phone, and its immediate context (previous and next phones). We tested the effect in the models of allowing a question about the language to which the phone belongs but we discarded it because it produced no noticeable improvement over models constructed with pure phonetic questions.

The percentage of states of the polyglot model that are shared by Spanish and Japanese phones depends on the total number of states. In models with 749 states, 45.5% of the states are shared, whereas in models with 9066 states, this percentage is 21.8%. For any given stop criterion, the number of states of the polyglot synthesizer is less than half the number of states of the two monolingual synthesizers together. Table 1 shows the percentage of shared and monolingual states for a 3-state triphone model with 2265 states.

Table 1 Percentage of Spanish, Japanese and mixed states

	Spanish	Japanese	Mixed
State 1	33.1%	22.2%	44.7%
State 2	32.7%	31.6%	35.7%
State 3	28.5%	26.1%	45.4%
TOTAL	31.6%	27.8%	40.6%

3.2 Speaker adaptation

When several speakers' voices are combined into a speaker independent model, the resulting voice is impersonal. Moreover, if the phone coverage of the training texts is not the same for all speakers, the voice identity can change unexpectedly in the middle of an utterance. In our system we are mixing not only speakers but also languages, therefore this effect is very likely. To improve the coherence of the output voice, we adapt the speaker independent voice to a specific speaker by means of supervised MLLR adaptation [7]. We decided to adapt only the mean values because the adaptation of the variances often produced unnatural values.

Generally speaking, the similarity to the original speaker increases with the number of adaptation matrices. However, when the language to be synthesized and the language of the target speaker's data are different, an excessive number of adaptation matrices degrade the basic quality of the synthetic

speech. To find out the optimum trade-off between similarity to the target speaker and speech quality under this cross-lingual condition, we adapted each model (polyglot, Spanish monolingual and Japanese monolingual) with 1, 4, 16, 64, 128 and 256 adaptation matrices, and pre-selected the adapted models with the best trade-off.

The polyglot model was adapted directly to Spanish and Japanese speakers. For cross-lingual speaker adaptation of the monolingual models we used phone-mapping. To adapt the speaker independent Japanese monolingual model to a Spanish voice, we mapped the phonetic transcriptions of the adaptation data onto Japanese phones. This mapping was done by rules. Basically we mapped each Spanish phone onto the phonetically closest Japanese phone. In some cases we used the results of the clustering tree of the bilingual model, e.g. we mapped Spanish [r] onto Japanese [g] instead of [ɾ]. We adapted the Spanish monolingual model to Japanese speakers in a similar way.

3.3 Synthesis

To synthesize speech, the phonetic transcription of a text is converted into a sequence of HMM states. For the polyglot synthesizer this conversion can be done directly for both languages because it has HMM models for Japanese and Spanish phones. In the case of the monolingual models, foreign phones have to be mapped onto their closest native ones. The mapping rules that we have used for synthesizing Japanese with the Spanish monolingual model are the same as for adapting the speaker independent Japanese model to a Spanish speaker.

4. EXPERIMENTAL CONDITIONS

To evaluate the performance of the polyglot model against the monolingual models we have performed a subjective evaluation of the understandability and the level of similarity between the adapted voice and the target speaker's voice.

4.1 Characteristics of the phonetic sets

The Spanish phonetic set is formed by 40 phones, including long and short vowels and the five diphthongs. The Japanese phonetic set is formed by 31 phones, including long and short vowels. Palatalized consonants have been modeled by the corresponding consonant followed by the semivowel [j]. The total number of different phones for both languages together is 54.

4.2 Characteristics of the HMM models

We used 3-states left-to-right triphone models without skips. Each state was modeled by 4 Gaussians. The transitions between states were modeled by state transition matrices.

The training data consist of 114 minutes of utterances from 10 Japanese male speakers and 104 minutes from 10 Spanish male speakers. All the data belong to the Globalphone corpus [12]. The polyglot model was trained with the 214 minutes of utterances from the 20 speakers.

The data 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.

For the monolingual sets, we pre-selected the models clustered with the MDL criterion [13], and for the polyglot set, the model with 2200 states (clustering threshold of 250). The number of states of the polyglot set was approximately the same as the number of states of the two monolingual sets together.

4.2 Model adaptation

Each speaker independent set was adapted to two Japanese and two Spanish male speakers. The adaptation data for each speaker was approximately 10 minutes.

The best trade off between basic quality and similarity to the target speaker was obtained with 16 matrices for the polyglot model and 4 matrices for the monolingual models.

4.3 Synthesis of the evaluation files

We synthesized 11 Japanese utterances using each one of the 12 adapted models. The length of each utterance was about 8 sec.

To focus only on the cepstral information, we used original prosody extracted from the audio files. The duration was extracted by a forced alignment of the evaluation files. The pitch was extracted with the ESPS function "get_f0" and synchronized to the phonetic segmentation. To adapt the pitch values to the characteristics of the target voices, we converted it into a logarithmic scale, normalized and adapted the mean pitch and average pitch range to the target voice.

4.4 Evaluation method

We evaluated the subjective intelligibility and the similarity between the synthetic voice and the original voice.

To evaluate the subjective intelligibility, 8 subjects were asked to score the evaluation utterances in a 5 points scale (1-very poor, 5-very good). The subjects were male native Japanese speakers with no hearing impairment or knowledge of Spanish. For each adapted models, we presented 3 files. For each target voice we presented 3 additional files, generated with the original cepstrum of the evaluation files and the original pitch adapted to the target voice. The utterances were presented randomly

Similarity to the target voice was evaluated by 8 subjects, with the same set of stimuli as for the subjective intelligibility plus 3 extra vocoder reconstructions of samples of the target speaker. The evaluation was performed for one voice at a time. Samples corresponding to the same voice were presented randomly. For subjects to get used to the target voice, we presented them a vocoder reconstruction of the original speaker's voice of around 10 seconds. This reference was played three times at the beginning and repeated immediately before every sample to be scored.

5. RESULTS

Figure 2 shows the results of the subjective intelligibility test. When adapted to Spanish voices, the polyglot model outperforms the monolingual models by almost one point. In the case of adaptation to Japanese voices the polyglot model presents the same score as the Japanese monolingual model. As expected, the subjective intelligibility of the Japanese model is significantly better when adapted to Japanese voices than to Spanish voices. Surprisingly, there is not such difference for the Spanish model in spite of the double phone-mapping for adaptation and synthesis

Figure 3 shows the results of the evaluation of the similarity. There is no significant difference among the three methods when they are adapted to Spanish voices in synthesizing Japanese. When adapted to a Japanese voice, the polyglot model and the Japanese monolingual model are also equivalent. The adaptation of the Spanish monolingual model to Japanese voices performs significantly worse than to Spanish voices.

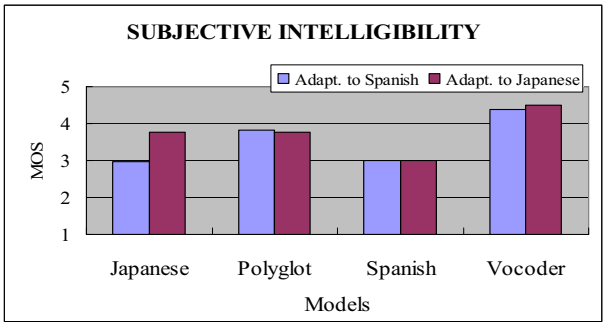

Fig. 2 Mean subjective intelligibility of Japanese Speech

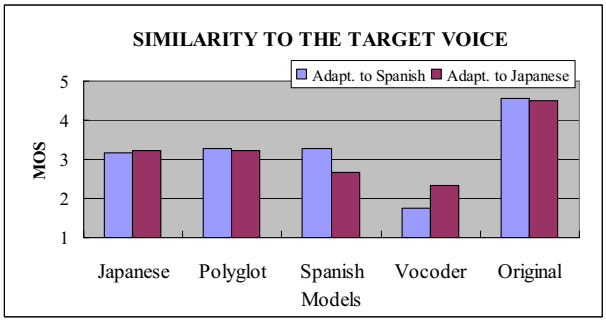

Fig. 3 Similarity to target speaker

This is due to phone mapping and the evaluation method. When comparing voices in two different languages, subjects concentrated only on the characteristics of the voice. Other features such as prosody or accent are assumed to be different and therefore not considered. However, for the Spanish monolingual model adapted to Japanese voices, subjects have to compare Japanese spoken by a native Japanese voice with Japanese spoken by a Spanish accented voice. Hence, unless voices are clearly similar, subjects judge them as different.

6. CONCLUSIONS

We have proposed a new method for multilingual synthesis that uses monolingual corpora to create a polyglot voice. This polyglot voice can be adapted to speakers of any of the languages included in the training data. The adapted models can synthesize speech with the same voice for all the languages of the training corpora.

In the case of adaptation to Spanish speakers and synthesis of Japanese texts, the subjective intelligibility of the proposed method outperforms those methods based on monolingual systems with phone-mapping for cross-lingual adaptation or synthesis. For Japanese target speakers, the subjective intelligibility of the proposed polyglot synthesizer is as good as a Japanese monolingual one. Informal tests suggest similar results could be achieved for Spanish synthesis.

Experimental results show that the similarity between the synthetic voice and the target voice is basically the same for the monolingual and the proposed polyglot method.

Since the size of the proposed method is smaller than the combined size of the two monolingual synthesizers, this method is desirable for multilingual applications that require minimal footprint.

7. FUTURE WORK

Our goal is to improve the quality of the synthetic speech, the similarity to the original voice, and increase the number of languages that can be synthesized.

In addition to integrating new languages into the training corpus, we want to evaluate the performance of the polyglot synthesizer when synthesizing and/or adapting to languages not included in the training data. This would be the case of minority languages. To achieve this, we are studying different methods to improve and automate the phone-mapping.

We want to record a bilingual or pseudo-bilingual corpus. Including such a corpus in the training data should increase the cohesion of the polyglot voice. It should also be useful to evaluate cross-lingual adaptation.

8. ACKNOWLEDGMENTS

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 their help and many helpful discussions.

This work is partly supported by the 21st Century COE-LKR Program.

9. REFERENCES

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

[2] J. Dijkstra et al., "Frisian TTS, an example of bootstrapping TTS for minority languages," *5th ISCA Speech Synthesis Workshop*, pp. 97-102, Jun. 2004

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

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

[5] L. Badino et al., "Language independent phoneme mapping for foreign TTS," *5th ISCA Speech Synthesis Workshop*, pp 217-218, Jun. 2004

[6] A. Black et al., "Multilingual text-to-speech synthesis," *ICASSP04* pp. 761-764, Mar. 2004

[7] 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

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

[9] J.L. Hieronymus, "ASCII phonetic symbols for the world's languages: Worldbet," *Journal of the International Phonetic Association*, 1993

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

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

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

[13] K. Shinoda et al., "MDL-based context-dependent subword modeling for speech recognition," *J. Acoustic Soc*, pp. 79-86, Mar. 2000