Cross-language Synthesis with a 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 examine the use of an HMM-based polyglot synthesizer for languages for which very limited or no speech data is available. In a former study, we presented a system that combines monolingual corpora from several languages to create a polyglot synthesizer. With this synthesizer we can synthesize any of the languages included in the training data with the same output voice and speech quality. In this paper, we approximate the sounds of non-included languages, by those available in the polyglot training data. Since the phonetic inventory of a polyglot synthesizer is wider than that of a monolingual one, the approximation of such non-included sounds becomes more accurate and thus the perceptual intelligibility increases. Moreover, the performance of a polyglot synthesizer can be further improved by adding a reduced amount of data from the target language.

1. Introduction

To develop a speech synthesizer in a new language is still a substantial task. In many cases, it requires large investments that nowadays are only profitable for a dozen or so languages. A possible solution to reduce the implementation costs is to reutilize speech resources from other languages. Most proposals in this direction are based on a phone mapping, which approximates the sounds of the target language by those of a similar language with an available speech corpus, e.g. [1].

Another possible solution is to use a polyglot synthesizer [2]. The wider "palette" of sounds available in a polyglot synthesizer with respect to a monolingual one, can make it easier to find appropriate candidates for the sounds of the target language. In this way, the approximated sounds can be closer to the real ones and the intelligibility of the synthesized speech increased.

In [3], we proposed a new approach to polyglot synthesis consisting in training a language independent HMM-based synthesizer with speech resources from several languages. For cross-language speech recognition, it was shown that a multilingual recognizer built in this way can outperform even the best-matched language dependent recognizer [4]. Moreover, if a small amount of data from the target language becomes available, it can be used to improve the performance of such HMM-based polyglot synthesizer. This can be done by adapting with it the polyglot synthesizer to the new language [5], or by including this new data in the training of the polyglot synthesizer.

2. HMM-based polyglot speech synthesis

A polyglot synthesizer is a system that can generate speech in different languages with the same voice. The two main approaches were: a) to record a corpus from a polyglot speaker [2] or b) to make a phonetic mapping between the phones of the language we want to synthesize and the phones available in the database [6]. In [3] we proposed a new approach that consists in combining monolingual corpora from several speakers in different language to train a language independent and speaker independent HMM-based synthesizer. The central assumption of our approach is that the average voice created by mixing data from several speakers tends to be language independent and therefore it can be considered as a polyglot voice. Figure 1 shows the general schema of our system. Since in our method no human polyglot talent is required we can expand it to any number of languages we want. Furthermore, since no phone mapping is needed for the languages included in the mixture, the perceptual intelligibility and the level of foreign accent when synthesizing these languages is lower than with other methods based on phone mapping.

The problem of synthesizing speech from an average voice is that it usually sounds impersonal. Moreover, there can be a lack of coherence in the resulting output voice, because not all the models are trained with data from the same speakers. To solve these two problems, we apply supervised Maximum Likelihood Linear Regression (MLLR) to adapt the average voice to the voice of a target speaker. Finally, we apply a synthesis algorithm [7] to the adapted HMM to generate speech in any of the training languages, independent of the language spoken by the target speaker.

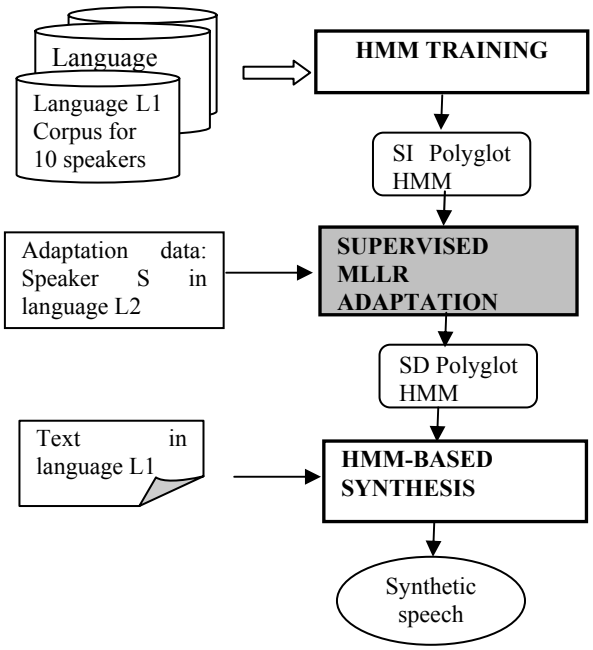

Figure 1: General schema of an HMM-based polyglot synthesizer.

The perceptual intelligibility obtained with a polyglot synthesizer when the language of the target speaker is different to language being synthesized, is significantly better than the perceptual intelligibility obtained by methods based on phone mapping. When the language of the target speaker and the language to be synthesized are the same, the polyglot synthesizer produces speech with the same perceptual intelligibility and similarity to the target speaker as a monolingual HMM-based synthesizer in that language.

3. Phone mapping

Although with an HMM-based polyglot synthesizer we can synthesize all the languages of its training data with the same voice and quality, if we want to synthesize text or adapt the average voice to speakers in other languages, we need to approximate the sounds of the target languages by those available in the training corpus. The easiest way to do this is phone mapping.

Table 1: Phone mapping applied to the Spanish phones.

Spanish	Icelandic	Japanese	Japanese+ Icelandic
Common to the 3 languages: f, i, j, k, m, n, o, p, s, t, w			
b	p	b (Eq)	b (Eq)
d	ð (Eq)	d (Eq)	d (Eq)
g	ɣ (Eq)	g (Eq)	g (Eq)
ŋ	n (Al)	ŋ (Eq)	ŋ (Eq)
tʃ	t + s	tʃ (Eq)	tʃ (Eq)
z	s (Al)	z (Eq)	z (Eq)
dʒ	j	dʒ (Eq)	dʒ (Eq)
θ	s (Di)	s (Di)	s (Di)
β	v	b (Al)	b (Al)
ɾr	r	ɻ	ɻ
x	h (Di)	h (Di)	h (Di)
ð	ð (Eq)	d (Al)	ð (Eq)
ɣ	ɣ (Eq)	g (Al)	ɣ (Eq)
a	a (Eq)	ɑ	ɑ (Eq)
l	l (Eq)	ɻ	l (Eq)
r	r (Eq)	ɻ	r (Eq)
u	u (Eq)	ɯ	u (Eq)
e	ɛ	e (Eq)	e (Eq)

Most approaches to phone mapping use the similarity between the articulatory features of source and target phones. These features are usually derived from the IPA representation of the phones, so that two sounds get associated if they share the same IPA symbol. The difficulties appear when no phone in the database shares exactly the same IPA representation as the target one and instead, there are several candidates that share the same number of articulatory features. To solve this problem, two possible approaches are a) asking a linguistic expert to build an ad-hoc assignment table for the target language [6] and b) assigning to each articulatory feature a language independent weight derived from perceptual tests [8]. Although the second approach is more attractive, it is debatable whether the same set of weights can be used for all languages. Many languages present allophonic and regional variations that modify the perceptual limits of their phonemes. These variations are specific for each language and very hard to predict. On the other hand, they can be exploited to improve the plausibility of the phone mapping. Many Spanish dialects for example, assimilate the phone [θ] to phone [s], or the phone [ʎ] to [j], [ʒ] or [dʒ], even though these phones belong to different phonemes in standard Spanish. For practical uses this means that to map [θ] to [s] or [ʎ] to any of its variants is an intra-lingual mapping, and consequently more acceptable for Spanish native speakers.

In our experiment we have used ad-hoc rules to map Spanish, to Icelandic, Japanese, and the phonetic set created by the addition of these two languages. In all the cases where it was possible, we have mapped the Spanish phones to a dialectal or allophonic variant of the phonemes to which they belong. Otherwise, we have assigned the Spanish phones to the phones with fewer articulatory differences and which produced the better perceptual intelligibility when synthesized. Table 1 shows the Spanish phones and their mapping into Icelandic, Japanese, and the bilingual phonetic set formed by Japanese plus Icelandic. The words between brackets indicates "equal phone" (Eq), "allophonic variant" (Al), and "dialectal variant" (Di). To map the affricate phone [tʃ] into Icelandic, we have used two phones, [t] and [s].

4. Experiments

The purpose of our experiments was to compare the perceptual intelligibility and level of foreign accent of the Spanish synthesized with Icelandic and Japanese monolingual systems versus a polyglot system trained with these two languages.

Whereas perceptual intelligibility scores how easily a subject understands the synthesized text, the level of foreign accent scores how plausible it sounds to a native speaker. This provides a measure of the naturalness of the synthesized speech.

To evaluate the performance of the polyglot synthesizer, we have trained a polyglot model that combines Japanese and Icelandic, "J+Ic", and two speaker independent monolingual models, one for each language.

To test how much we can improve the polyglot synthesizer by adding speech data in the target language, we have created two additional trilingual models. In the first one, we have added 20 minutes of Spanish speech to the Japanese and Icelandic data. This represents 20% of the data used for the other two languages. In the second model, we have added the same amount of Spanish speech as for the other two languages: around 100 minutes. We will refer to these two models as "J+Ic+0.2S" and "J+Ic+S" respectively.

We have adapted each model with supervised MLLR to one Japanese and one Icelandic speaker not included in the training data. The monolingual models were adapted only to the speaker in their corresponding language and the polyglot ones to both speakers.

4.1. Training and adaptation data

The training data for the Japanese monolingual model consists of ten male speakers of the Globalphone corpus [9]. We selected the speakers with more data available and whose voices seemed to us more similar one another.

The Icelandic model was also trained with data from 10 male speakers selected with the same criteria as for Japanese. The data in this case belong to the Jensson's corpus [10].

The "J+Ic" model was trained with the same data as the two monolingual models.

The Spanish data added to the trilingual models belong to the Globalphone corpus. For "J+Ic+0.2S", we have added the data of two Costa Rican speakers and to "J+Ic+S" those of nine Costa Rican speakers and one Mexican.

The speech data for each speaker was approximately 10 minutes. For adaptation, we have also used 10 minutes of data for each target speaker.

We should mention that neither the Globalphone nor Jensson's corpora were designed for speech synthesis.

4.2. Transcription labels

The labels are a modified version of IPA. For some compound phones such as diphthongs and palatalized, we have preferred to use two separated phones.

In these experiments we assume that two sounds that share the same IPA symbol are similar enough to be represented by the same HMM.

4.3. HMM models

The models are triphone HMMs with 1 Gaussian, 3 states left-to-right without skips.

The feature vector consists of 25 mel-cepstral coefficients and their delta, calculated from a 16 ms Blackman window with a 5 ms shift. We use a short analysis window due to the labeling of diphthongs and palatalized phones with two triphones. Otherwise the minimal duration required by the state sequence of 6 states could become longer than the real duration of the sounds.

To adapt the speaker independent models we have applied unconstrained supervised MLLR adaptation to the mean value of the pdfs. For all the models we have used 4 adaptation classes.

To cluster the "J+Ic" model with a phonetic decision tree, we have chosen a ML threshold that produced a similar number of final leaves as for the Japanese and Icelandic monolingual models.

4.4. Prosody

In order to consider only the effects on the intelligibility of the phone mapping, we have preferred to use original prosody. The duration corresponds to a Viterbi forced alignment of the Spanish texts, and the pitch was automatically extracted and synchronized to each sequence of mapped phones.

4.5. Evaluation method

For the evaluation, we have asked 5 native Spanish speakers (3 Mexican, 1 Bolivian and 1 Spaniard) to evaluate using a 5 point MOS scale the perceptual intelligibility and the level of foreign accent of 24 samples each. For the foreign accent, 5 points MOS means "native speaker".

We have evaluated 30 different texts synthesized by the 4 different models: Monolingual, "J+Ic", "J+Ic+0.2S" and "J+Ic+S", each one adapted to a Japanese and an Icelandic speaker. The samples were selected in such a way that every combination of text and model were evaluated once, and each subject listened to each text at most twice. The samples were presented to the subjects pseudo-randomly trying to separate as much as possible the repetition of the same text. The whole test took around 30 minutes for each subject.

5. Results

Figure 2 shows the perceptual intelligibility and foreign accent of the models adapted to an Icelandic speaker. It can be appreciated that the perceptual intelligibility increases with the addition of new languages. However, the system does not improve just by adding more data from the target language. There is no significant difference between the understandability of the "J+Ic+0.2S" model with only 20 minutes of Spanish data, and that of the "J+Ic+S" model with 100 minutes.

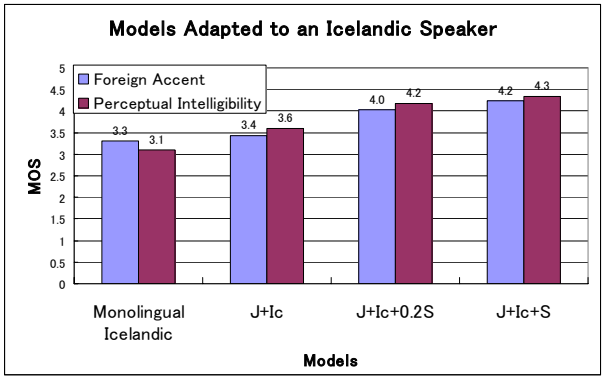

Figure 2: Results of the models adapted to an Icelandic voice.

Figure 3 shows the results of the models adapted to a Japanese speaker. In this case, the addition of Icelandic to the monolingual model does not produce any significant change of the perceptual intelligibility. With 20 additional minutes of Spanish, there is a slight improvement of the perceptual intelligibility but not yet significant. The real effect of this extra data is a reduction of the foreign accent. We need to add the full amount of Spanish data to obtain a statistically significant improvement.

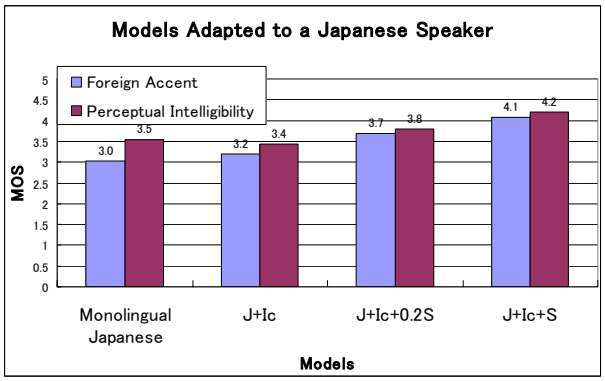

Figure 3: Results of the models adapted to a Japanese voice.

The perceptual intelligibility of the "J+Ic" and "J+Ic+S" models adapted to Icelandic and Japanese voices is the same but for the "J+Ic+0.2S" model it is better when adapted to an Icelandic speaker.

The foreign accent of the monolingual and "J+Ic" models, does not present any significant difference either for Japanese

or for Icelandic voices. The foreign accent only improves when Spanish data is added and it is the same with 20 or 100 minutes of Spanish data for Icelandic or Japanese voices.

6. Discussion

The perceptual intelligibility achievable by means of phone mapping depends on the distance between the target and source languages. However, the proportion of Spanish phones mapped into a phone with a different IPA representation is 30% for both Japanese and Icelandic. In order to explain the results of the evaluation, we need to redefine the distance between languages.

If we consider a language as a set of phones Set_a, we can define the mean phonetic distance *Dist* between two phonetic sets Set_a and Set_b as:

$$Dist(Set_a, Set_b) = \sum_{ph \in Set_a} d(ph, Set_b) \cdot P(ph, Set_a) \quad (1)$$

where $d(ph,Set_b)$ is the distance between phone *ph* and the phones of Set_b., and $P(ph,Set_a)$ is the occurrence probability of phoneme *ph* in the Set_a. For $d(ph,Set_b)$ we have used the following definition:

$$d(ph, Set_b) = SymKL(ph, ph^{map}) \quad (2)$$

where ph^{map} is the phone of Set_b to which *ph* is mapped, and $SymKL(ph,ph^{map})$ is the mean symmetric Kullback-Leibler distance between the monophone HMMs of phones *ph* and ph^{map}.

Table 2 shows the mean phonetic distance between the Spanish phonetic set and the Icelandic, Japanese and "Japanese+Icelandic" phonetic sets when the phone mapping defined in Table 1 is used.

Table 2: Phonetic distance between Spanish and its mapping into Japanese, Icelandic and J+Ic.

Phonetic set	**Mean phonetic**
Japanese	16.2
Icelandic	28.6
Japanese+Icelandic	16.3

As we can see, for Icelandic the addition of the Japanese phones reduces the phonetic distance by 43% absolute. However for Japanese, the phonetic distance is basically the same with or without the Icelandic phones. This explains the results obtained for the perceptual intelligibility of the Japanese and Icelandic monolingual models and the "J+Ic" model.

7. Conclusions

In this paper we have examined the performance of a polyglot synthesizer to synthesize languages for which very limited or no speech data is available.

The intelligibility of the speech synthesized using phone mapping depends on the phonetic similarity between the target and source language. However, to calculate this similarity a priori can be quite complicated. By using a polyglot synthesizer we can ignore this problem. The perceptual intelligibility of a polyglot synthesizer is as good as the perceptual intelligibility of the best monolingual synthesizer trained in any of its languages.

The perceptual intelligibility and foreign accent of a polyglot synthesizer can be improved by adding extra speech data of the target language. Depending on the language of the target speaker, adding just 20% of data from the target language can be equivalent to a full polyglot synthesizer that includes the target language.

8. Future work

Our next steps will be directed to develop an algorithm that instead of mapping the unseen phones to those available in the training data, interpolates the available phones to create new acoustic models. We want to evaluate this approach for different languages, using a polyglot synthesizer that includes four or five languages.

We also want to investigate the average amount of data from the target language needed to obtain a performance equivalent to a polyglot synthesizer that fully includes that language, and how it varies depending on the phonetic distribution of the target language.

9. Acknowledgements

This work was partially funded by the 21st Century COE-Large-scale Knowledge Resources Program.

10. References

[1] Dijkstra, J., Pols, L.C.W. and van Son, R.J.J.H., "Frisian TTS, an example of bootstrapping TTS for minority languages", *5th ISCA Speech Synthesis Workshop*, pp. 97-102, Pittsburgh, USA 2004.

[2] 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.

[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] Schultz, T. and Waibel, A., "Experiments on cross-language acoustic modeling", *Proc. Eurospeech*, pp. 2721-2724, Aalborg, Denmark 2001.

[5] Koehler, J. "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks", *Proc. ICASSP*, pp. 417-420, Seattle, USA 1998.

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

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

[8] Badino, L., Barolo, C. and Quazza, S., "Language independent phoneme mapping for foreign TTS", *5th ISCA Speech Synthesis Workshop*, pp. 217-218, Pittsburgh, USA 2004.

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

[10] Jensson, A.T., Whittaker, E.W.D., Iwano, K. and Furui, S., "Language model adaptation for resource deficient language using translated data", *Proc Interspeech*, Lisboa, Portugal 2005.