New approach to polyglot synthesis: how to speak any language with anyone's voice

Javier Latorre, Koji Iwano, Sadaoki Furui.

Department of Computer Science
Tokyo Institute of Technology, Tokyo, Japan
E-mail: {Latorre, iwano, furui}@furui.cs.titech.ac.jp

Abstract

In this paper we present a new method to synthesize multiple languages with the voice of any arbitrary speaker. We call this method "HMM-based speaker-adaptable polyglot synthesis". The idea consists in mixing data from several speakers in different languages to create a speaker-independent multilingual acoustic model. By means of MLLR, we can adapt this model to the voice of any given speaker. With the adapted model, it is possible to synthesize speech in any of the languages included in the training corpus with the voice of the target speaker, regardless of the language spoken by that speaker. When the language to be synthesized and the language of the target speaker are different, the performance of our method is better than that of other approaches based on monolingual models and phone mapping. Languages with no available speech resources can also be synthesized with a polyglot synthesizer by means of phone mapping. In this case, the performance of a polyglot synthesizer is better than that of any other monolingual synthesizers based on languages which were used to train the polyglot one.

1. Introduction

As a result of the globalization process, the usage of two or more languages has become a daily routine for an ever increasing number of people. Since the learning of a new language is a hard task, it is logical to suppose that those who have to use multiple languages will demand software applications that can deal with multiple languages too. Moreover, they will expect computers to assist them with those languages they cannot speak fluently.

An additional requirement of some speech synthesis applications is the ability to transform the output voice into the voice of any new speaker, without recording much speech data from that speaker. In a multilingual framework, e.g., a speech-to-speech translator, the language of the target speaker will often be different than the language we want to synthesize.

The goal of this research is the development of a system that can synthesize multiple languages with any arbitrary voice. With such a system, a user would be able to synthesize speech with his own voice in languages which he cannot speak himself. We call such a system a 'speaker adaptable polyglot synthesizer'.

Another application of a polyglot synthesizer is the synthesis of speech for minority languages. For languages with very limited or no available speech data, the most common way to implement a speech synthesizer is to use a synthesizer from another language and transform it to make it speak the new language. The most common transformation is the phone mapping between the target language and the language of the available synthesizer. For phonetically similar language this method works acceptably well. However, this method always introduces an error which varies with respect to the phonetic similarity between the target and the substitute language. By using a polyglot synthesizer we can reduce the mapping error and thus the quality of the synthesized speech can be improved. This allows us to create synthesizers for new language at low cost and with acceptable quality. This aspect of our approach makes it especially interesting for minority languages.

2. Polyglot synthesis

The two previous approaches to synthesis several languages with the same voice consisted in a) recording speech data from a real 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 a very difficult challenge. Moreover, the system cannot be expanded to languages other than those spoken by the polyglot voice talent.

The second approach [2] can be applied to any language and any speaker. However, the resulting voice has a very strong foreign accent which makes it difficult to understand. As a result, unless the target and substitute languages are phonetically close, the resulting voice becomes almost unintelligible.

3. HMM-based polyglot synthesis

In [3], we proposed a new method to create a polyglot synthesizer that consists in training an HMM-based synthesizer [4] with data from several monolingual speakers from each one of the languages we want to synthesize. Our assumption is that voice differences depend only on anatomical factors. Therefore the average voice created by mixing a sufficient number of speakers tends to be the same for any language.

The architecture of our proposal is shown in Fig. 1. It has three steps: training, adaptation and synthesis.

3.1. HMM Training

In the first step, a speaker- and language-independent acoustic HMMs (SI Model) is trained with speech data from several monolingual speakers in multiple languages.

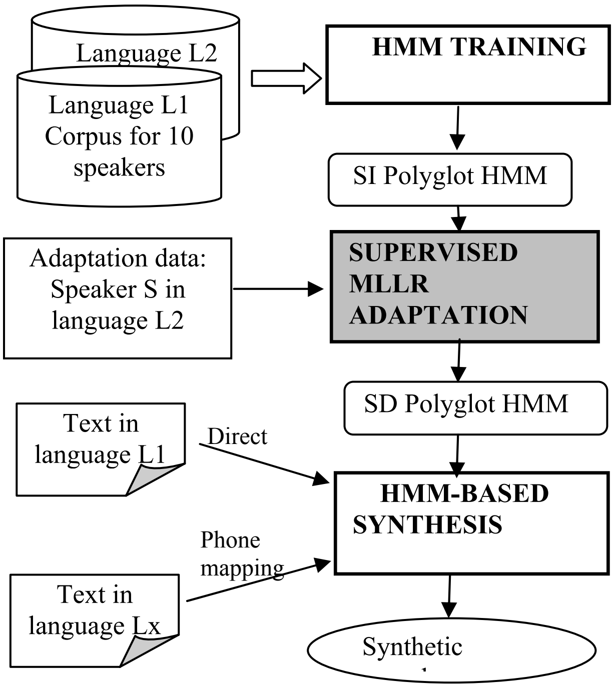

Figure 1: **Architecture of an HMM-based speaker adaptable polyglot synthesizer.**

In order to combine all the languages into a single acoustic model, the phonetic labels of each language are normalized to a common IPA code, so that the same label refers to the same IPA phone.

After training triphone models with the normalized labels, the resulting triphones are clustered using a phonetic decision tree. We use only one decision tree per state for all the phones. In this way, we can profit from the similarity across phones [5]. The questions of the decision trees refer only to the phonetic features of the phone and its immediate left and right context.

3.2. Supervised MLLR Speaker adaptation

In the second step, the SI-HMMs are adapted by means of MLLR [6] to the voice of a target speaker. The function of the adaptation is on the one hand to provide the output voice with an identity and on the other hand to assure that the synthesized voice sounds equal for all the languages. Since the phonetic coverage of the speakers used to train the SI model is not the same for each phone, different models were trained with data from different speakers. As a result, if no speaker adaptation is performed there might be a risk that the identity of the output voice changes in the middle of a sentence. By means of MLLR this problem is practically eliminated.

In general, the similarity to the target speaker increases with the number of adaptation classes. However, an excessive number of adaptation classes degrades the intelligibility of the synthesized speech. The optimum number of classes depends mainly on the amount of adaptation data and on the number of HMMs of the SI model to be adapted. For the evaluation of our system we planned to test the performance of 66 SD models made from the combination of different languages and adapted to several speakers. To evaluate every combination of model size and number of adaptation classes for each one of the 66 models would have required too much time and too many subjects. In order to avoid this problem, we pre-selected for the evaluation only those SD models that after a preliminary evaluation presented the better trade off between intelligibility and similarity to the target speakers.

3.3. HMM-based speech synthesis

In the third stage, the SD models are used to generate the speech waveform. The synthesis process has two parts. In the first one a sequence of Mel-Cepstral coefficients (MCC) is generated according to the sequence of input phones [4].

In the second part, the sequence of MCCs are used as parameters of a Mel Log Spectral Approximation (MLSA) filter [7]. The speech signal is generated by filtering with this MLSA filter a pulse-noise signal created according to the voiced/unvoiced characteristics of the phones.

3.4. Synthesis of extrinsic languages

In our approach, any language included in the training data can be synthesized directly. However to synthesize other languages, it is necessary to approximate the sounds of the new language by the sounds of the languages included in the model. In our system this approximation is done by means of an IPA-based phone mapping. The set of phones that is obtained when multiple languages are combined is obviously greater than the set of phones that can be obtained from a single language. Consequently, using a polyglot model, it is more likely to find good approximations to the sounds of the new language. By using more accurate approximations, the mapping error is decreased and thus the level of native accent and the intelligibility of the synthesized speech increase.

4. Experiments

By means of a perceptual evaluation we wanted to test the feasibility of our approach and compare it with other possible approaches based on phone mapping for synthesis or adaptation. We evaluated our system under three different scenarios:

-Cross-language synthesis, when the language to be synthesized and the language of the target speaker are different and both included in the training data.

-Synthesis of extrinsic languages, when the language of the target speaker is included in the training data but the language to be synthesized is not.

- Direct synthesis, when the language to be synthesized is the same as the language of the target speaker and it is included in the training data.

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

The performance of the different models was measured according to three parameters:

- Perceptual Intelligibility: how easy it is for subjects to understand the synthesized speech. This parameter is equivalent to speech quality but focuses only on intelligibility and ignores other factors such as naturalness, noise, etc.

- Similarity between the synthesized voice and the target speaker.

- Native accent: whether the synthesized speech sounds like a native speaker or a foreigner. This parameter is intended to give an idea about the naturalness of the synthesized voice.

4.1. Trained models

Using different combinations of Spanish, Japanese, German, French and Russian data, we trained several acoustic models. The models are tied-state triphone models with 1 Gaussian, 3 states, and left-to-right structure without skips. The feature vector consists of the total energy, the 25 first mel-cepstral coefficients and their delta. The analysis window is a 16ms Blackman window with a 5 ms shift.

Depending on the combination of languages used in the training, the SI model can be classified into three groups:

-Monolingual models.

-Polyglot models mixing Spanish and two other languages out of Russian, French, and German.

-Polyglot models mixing Japanese and two other languages out of Russian, French, and German.

Each model was trained with roughly the same amount of speech data from 30 monolingual male speakers. For the polyglot models the training data was distributed equally among the three languages, thus 10 speakers were used for each language included in the model. The models pre-selected for the evaluation according to the criterion mentioned in Section 3.2 were those models whose size was determined using the Maximum Description Length criterion [8]. Table 1 shows the size and amount of training data for each SI model.

Table 1 **Characteristics of the pre-selected models**

Language mixture	#phones	#final states	Training data (minutes)
Japanese (Ja)	26	2389	349.4
Russian (Ru)	30	2377	306.1
French (Fr)	34	3218	463.22
Spanish (Sp)	29	2410	316.41
German (Ge)	32	2648	334.01
Ja+Ge+Fr	47	4476	377
Ja+Ru+Fr	48	4526	331.75
Ja+Ru+Ge	43	4062	365.82
Sp+Ge+Fr	48	4280	380.68
Sp+Ru+Fr	48	4102	369.49
Sp+Ru+Ge	44	3709	335.43

4.2. Adaptation

After the training of the SI-HMMs, we used supervised MLLR to adapt them to the voices of several target speakers. Every model was adapted to two speakers of each language included in its training data. Additionally Spanish and Japanese monolingual models were also adapted to two speakers of each one of the other two languages not included in their training data. To perform the adaptation to these speakers, prior to the proper MLLR adaptation of the Spanish and Japanese monolingual SI models, the labels of the adaptation data were mapped into the respective Spanish and Japanese phone sets.

For speakers of languages included in the training data, models adapted with 4 adaptation classes were pre-selected. To adapt the Spanish and Japanese monolingual models to speakers of other languages, the models pre-selected were those adapted with 2 adaptation classes.

4.3. Training data

Training and adaptation data has been extracted from the Globalphone speech database[9]. The Globalphone database was not specifically designed for speech synthesis, however it was the best available database for fulfilling our need for multilingual data and multiple speakers for each language.

We selected for each language those speakers whose voices we found more similar one another.

4.4. Synthesis

The synthesis of Spanish or Japanese texts by models that included these languages was done directly. For the other models we used phone mapping. The mapping rules were defined based on the phonetic similarity between phones. Whenever possible, each Spanish and Japanese phone was substituted by phones with the same IPA representation. Otherwise, the substitute phones were those available phones with the highest phonetic similarity. In such case, when two or more phones showed the same similarity, we selected the phone that in the original language (Spanish or Japanese) is considered an allophonic variant of the phone to be mapped.

4.5. Experimental conditions

For the evaluation, 18 Spanish and 18 Japanese texts were synthesized by each one of the 66 adapted models. These files were presented to 6 native Japanese and 6 native Spanish subjects respectively. Due to the high number of total stimuli to be evaluated for each language (66 model x 18 texts = 1188 stimuli) we distributed them among the 6 subjects so that each subject listened to 3 test files from each one of the 66 adapted models. Still, the total number of stimuli per subject was quite large, therefore the stimuli assigned to each subject were distributed across 3 evaluation sessions of about one hour each. The distribution of the stimuli into sessions was done in such a way that at each session at least one stimulus for each model was presented and the same test texts were not listened more than 4 or 5 times. The stimuli of each session were presented in a random order. In addition to the synthesized stimuli, the vocoder re-synthesis of the audio version of the test texts was also included at each session. In this way the scores across sessions can be considered to be in the same scale.

The three evaluation parameters were evaluated on a 5 point MOS scale simultaneously. To evaluate the similarity to the target speaker, subjects were asked to compare the synthesized voice with a short reference audio file with the voice of the target speaker. This reference was presented before the synthesized file. In order to keep the consistency of the experiment, the reference file was also presented before the re-synthesized files.

4.6. Prosody

The purpose of this experiment was to focus on the acoustic models. Therefore to avoid the interference produced by different prosodic models, we decided to use original prosody extracted from the natural speech of the read version of the texts we used in the evaluation. In this way, the duration of the phones was estimated by means of a Viterbi forced alignment of the test texts, and the F0 was extracted using the ESPS function “get_f0”.

In order to approximate the original prosody to the prosody of the target speakers, we shifted the mean F0 of each test file to the mean F0 of each target speaker. No modification was applied to the duration.

5. Results

The following figures show the results of the evaluation for the three scenarios mentioned in section 4. The columns named “Monolng. phone mapping synth.” represent the average scores of monolingual models in the language of the target speakers that use phone mapping to synthesize the target language. The columns named “Monolng. phone mapping adapt.” represent the average of Spanish and Japanese models adapted to target speakers in other languages by means of phone mapping , i.e., cross-language adaptation, 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-language synthesis, synthesis of extrinsic languages, and direct synthesis respectively. The columns named “Vocoder” represent the scores obtained by the vocoder re-synthesis of the test texts. This column is the ceiling for perceptual intelligibility and native accent, and the noise level for similarity. Error bars indicate the standard deviation

The relative scores for the three evaluation parameters were similar for Japanese and Spanish. However, the scores of the perceptual intelligibility for Spanish were almost one point higher than their Japanese equivalents for all the models.

5.1. Cross-language synthesis

Figure 2 shows the performance of the different models in the case of cross-language synthesis. It can be seen that the perceptual intelligibility and native accent obtained by the “Polyglot cross” model is much better than that obtained by the monolingual models “Monolng. phone mapping adapt” or “Monolng. phone mapping synth”. The differences between the polyglot models and the “Monolng phone mapping adapt” model is not so great. Nonetheless, this difference is still significant in terms of perceptual intelligibility. In terms of native accent both models perform equally well. With respect to the perceived similarity to the target speaker, the three models were basically the same.

5.2. Synthesis of extrinsic languages

Figure 3 shows the average result in the case of extrinsic languages. It can be seen that the performance of the polyglot models clearly surpasses that of the monolingual ones when a new language has to be synthesized by means of phone-mapping. The performance of the polyglot models under this scenario was indeed better than the performance of any of the monolingual models.

As Figures 4 and 5 show, the perceptual intelligibility of the “Monolng. phone mapping synth” models, depends largely on the language they have been trained with.

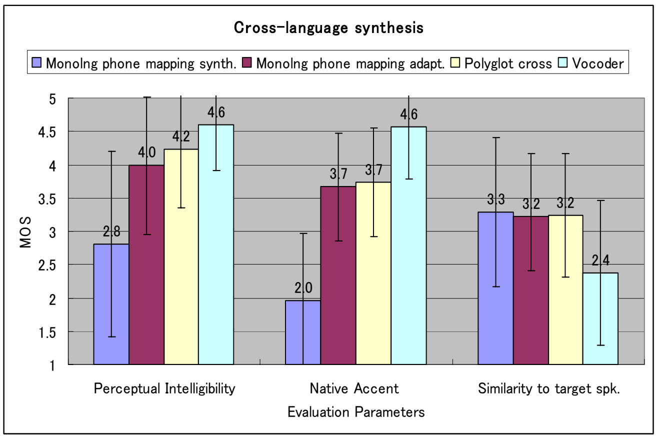

Figure 2: **Performance for cross-language synthesis.**

When the language of the model is phonetically close to the language to be synthesized, the results are acceptable; otherwise the synthetic voice is almost unintelligible. On the other hand, when we use a polyglot synthesizer, the results are almost the same independently of the language of the target speaker. No significant difference was found between models trained with different languages mixtures.

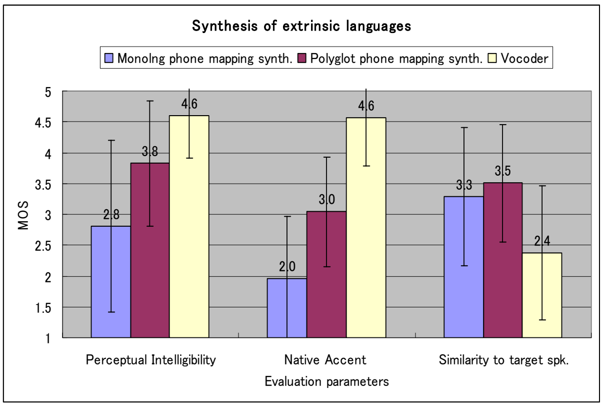

Figure 3: **Performance for the synthesis of extrinsic languages.**

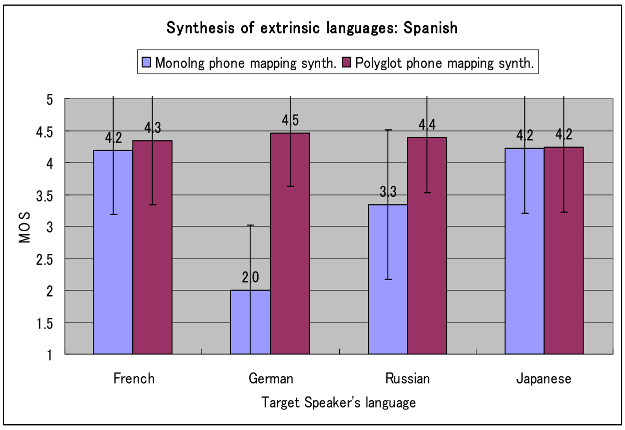

Figure 4: **Perceptual intelligibility for the synthesis of Spanish as a extrinsic language.**

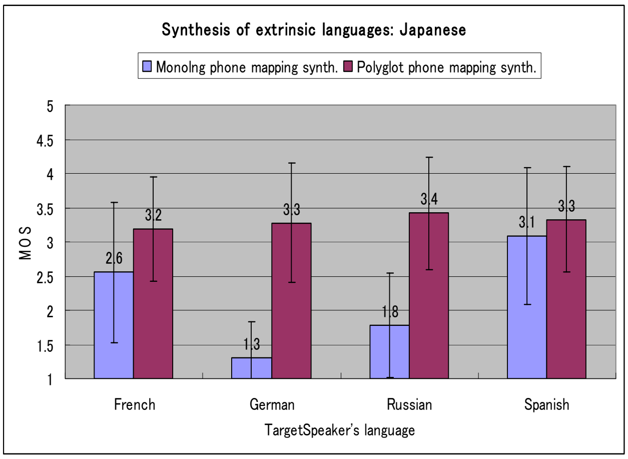

Figure 5: **Perceptual intelligibility for the synthesis of Japanese as a extrinsic language.**

5.3. Direct synthesis

Direct synthesis is the scenario under which synthesis is most commonly carried out. We included this scenario as a control to estimate the degree for which mixing languages degrades the performance with respect to a standard monolingual synthesizer.

Figure 6 shows the results for direct synthesis. It can be seen that the differences between the "Polyglot Direct" model and the "Monolng. Direct" models are rather small. The scores for the polyglot model in terms of perceptual intelligibility are slightly worse than that of the monolingual models but this difference was not statistically significant. In term of the other two parameters the polyglot model again yielded performance levels which were just slightly worse than those of the monolingual models.

Figure 6 also depicts the scores of the "Polyglot cross" models. It can be seen that although the performance of the polyglot models in cross-language synthesis was worse than in direct synthesis, the difference in terms of perceptual intelligibility is not that great.

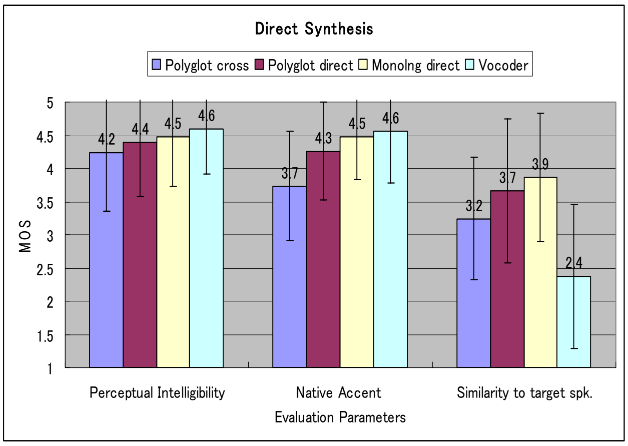

Figure 6: **Performance for Direct Synthesis.**

6. Discussion

6.1. Language similarity vs. perceptual intelligibility

As it can be seen in Figures 4 and 5, the perceptual intelligibility in the synthesis of extrinsic languages depends on the language of the model. These differences of the perceptual intelligibility can be explained by the acoustic distance between the target and substitute language or language combination. Given a phone mapping function ϕ *a,b,* between two languages Lng_a and Lng_b, the acoustic distance between them can be defined as the mean acoustic distance of the corresponding monophone models. That is

$$Dist(Lng_a, Lng_b) = \sum_{ph \in Lang_a} d(ph, Lng_b, \varphi_{a,b}) \cdot P(ph, Lng_a) \qquad (1)$$

where *d(ph,Lngb,* ϕ *a,b)* is the distance between phone *ph* of *Lnga* and the corresponding phone of *Lngb* assigned by ϕ *a,b*, and *P(phLnga)* is the occurrence probability of *ph* in *Lnga*. A possible measurement of this acoustic distance is the mean symmetric Kullback-Leibler divergence, defined as:

$$d(ph, Lng_b, \varphi_{a,b}) = \frac{1}{S} \sum_{s=1}^{S} SymKL(ph^s, \varphi_{a,b}(ph)^s) \qquad (2)$$

where ph^s is the *s* state of the acoustic mophone HMM of *ph.*

In Figure 7 we can see that the relationship between perceptual intelligibility and acoustic distance as defined in (1) is almost linear. The correlation coefficient for the Japanese evaluation was > 0.99 and for the Spanish evaluation > 0.97. This relationship explains the difference in terms of perceptual intelligibility that we have found between the polyglot and monolingual models. Since the acoustic distance of the "Polyglot phone mapping synth" model is less than or equal to that of any of the languages used to train the model, the perceptual intelligibility is also higher than or equal to that of the monolingual model in the language closest to the target language.

It is interesting to note that although the absolute level of the perceptual intelligibility for Japanese and Spanish was different, the slope with respect to the acoustic distance is practically the same. If Japanese and Spanish scores are set a the same level by subtracting the mean difference of 0.97 MOS points, the correlation factor between the perceptual intelligibility and the acoustic distance is >0.98.

6.2. Perceptual intelligibility vs. Native accent

The results obtained for the evaluation parameter 'native accent' are very similar to those of the 'perceptual intelligibility'. Indeed, we found that these two parameters have a strong statistical interdependence. However, unlike the 'perceptual intelligibility', the 'native accent' does not depend on the language of the monolingual model We have not found any significant difference between the native accent either for monolingual models in different languages or for polyglot models with different language mixtures.

One possible explanation for the better 'native accent' results obtained by the polyglot models against the monolingual models in the scenario of synthesis of extrinsic languages is that in the case of monolingual models, subjects listen to voices with an identifiable accent, proper of speakers of that language. However, in the case of the polyglot models the accent of the synthetic voice is mixed and more difficult to identify. Therefore subjects cannot decide so clearly whether the synthesized voice sounds like a foreigner or not.

7. Conclusions

In this paper we have presented our approach to polyglot synthesis applied to two languages: Japanese and Spanish. We have shown that it is possible to build a polyglot synthesizer by mixing data of multiple monolingual speakers in different languages. Furthermore, since our approach is based on HMM-synthesis, the system can be easily adapted to imitate the voice of any given speaker.

With our approach, we can synthesize with the same voice all the languages included in the training data with similar quality. In the case of direct synthesis, this quality does not differ significantly from that obtained with a monolingual model.

We have shown that for cross-language synthesis our system performs much better than an approach based on monolingual acoustic models and phone mapping for synthesis.

Although the performance of the approach based on cross-language adaptation of a monolingual model was not much worse than that of the polyglot models to synthesize multiple languages by means of cross-language adaptation, we need to adapt as many monolingual synthesizers as languages we want to synthesize. This implies that the adaptation of each monolingual synthesizer is done independently increasing the risk that the output voice might not be the same for all the languages. On the contrary, in a polyglot synthesizer all the languages are adapted together therefore this risk is much lower.

Languages not included in the training data of the polyglot model, were also synthesized by means of phone-mapping. In this case, the performance levels obtained by the polyglot models were better than those obtained by the monolingual models trained in any of the languages mixed to create the polyglot model. This result makes our approach especially attractive for minority languages, for which the amount of available speech data is usually very limited or inexistent.

8. Future work

In the future we plan to investigate different methods to implementing the phone mapping function.

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

Finally we want to investigate how many data from a new language should be added to a polyglot synthesizer in order for the synthesizer to obtain a performance for that language equivalent to the one obtained for the other languages already included in the system.

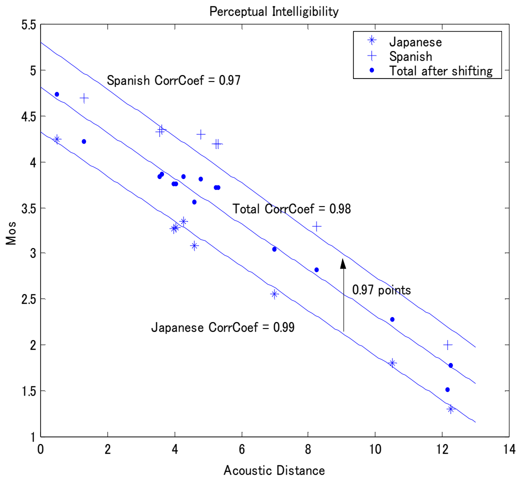

Figure 7: **Acoustic distance vs. Perceptual Intelligibility**

Acknowledgements

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

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

[7] Imai S., "Cepstral analysis synthesis on the mel frequency scale", *Proc. ICASP*, pp. 93-96, Boston, USA 1983.

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

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