Towards Cross-Lingual Emotion Transplantation

Jaime Lorenzo-Trueba[1], Roberto Barra-Chicote[1],
Junichi Yamagishi[2], and Juan M. Montero[1]

[1] Speech Technology Group, ETSI Telecomunicacion,
Universidad Politecnica de Madrid, Spain
[2] National Institute of Informatics, Tokyo, Japan
{jaime.lorenzo,barra,juancho}@die.upm.es
jyamagish@nii.ac.jp

Abstract. In this paper we introduce the idea of cross-lingual emotion transplantation. The aim is to lean the nuances of emotional speech in a source language for which we have enough data to adapt an acceptable quality emotional model by means of CSMAPLR adaptation, and then convert the adaptation function so it can be applied to a target language in a different target speaker while maintaining the speaker identity but adding emotional information. The conversion between languages is done at state level by measuring the KLD distance between the Gaussian distributions of all the states and linking the closest ones. Finally, as the cross-lingual transplantation of spectral emotions (mainly anger) was found out to introduce significant amounts of spectral noise, we show the results of applying three different techniques related to adaptation parameters that can be used to reduce the noise. The results are measured in an objective fashion by means of a bi-dimensional PCA projection of the KLD distances between the considered models (neutral models of both languages, reference emotion for both languages and transplanted emotional model for the target language).

Keywords: Statistical Parametric Speech Synthesis, Expressive Speech Synthesis, Emotion Transplantation, Cross-lingual.

1 Introduction

In nowadays society we can see that computers are increasingly present: internet, smart phones, tablets or virtual agents are just a few examples of machines we have included in our daily life. In consequence, numerous fields of study around these machines have appeared. Among them, the study of human-machine interactions and interfaces pose a challenge, as providing simple and efficient communication interfaces could significantly reduce the gap in technology usability.

Speech synthesis is one of the most natural ways of providing such human-machines interfaces. The objective of speech synthesis systems is to produce artificial human speech by means of either hardware or software technologies. Nowadays speech synthesis can provide very good quality when producing neutral speech regardless of the technology [2] which is ideal for speech interfaces

J.L. Navarro Mesa et al. (Eds.): IberSPEECH 2014, LNAI 8854, pp. 199–208, 2014.

that do not need to engage in a conversation with the user. On the other hand, applications where simulating a more natural behavior is necessary, imbuing the synthetic speech with expressive features (e.g. emotions, speaking styles...) becomes very interesting. This is the role of expressive speech synthesis.

Expressive speech synthesis presents significant challenges: maintaining a good speech quality with traditional systems can be problematic because of how variable some expressiveness are, recording good quality stable expressive data is also difficult because it is difficult for non professionals to maintain stable speaking patterns even while speaking normally, and also as there are so many possible expressiveness that can be produced when talking in real life, it is nearly impossible to cover all of them. In the end, one of the biggest problems is data acquisition. The work proposed in this paper aims to fix one of the main shortcomings of expressive speech synthesis: scalability. We want to obtain a method capable of learning the expressive nuances of emotional speech, and transplant it to different speakers for whom we do not have any expressive information across languages.

One approach to emotion transplantation is the use of projective adaptation techniques such as Cluster Adaptive Training (CAT) [3], Eigenvoices [9] or Multiple-Regression Hidden Semi-Markov Models (MR-HSMM) [6]. These techniques are capable of imbuing emotions into the target adaptation speaker as long as the emotional data was included in the original training process because the output is always a combination of the training models, which at the same time makes them extremely robust and capable of providing very high speech quality transplanted models. As a shortcoming, the output can only take a limited amount of values, so speaker similarity cannot be guaranteed as the transplantation reach is constrained. Another approach is the use of rules to modify the features of the speaker models. This approach should theoretically be capable of modifying any neutral speaker model so that it conveys the desired emotion as long as the correct rules are applied. The truth is that these approaches are tipically capable of correctly imbuing the desired emotion and even provide reasonably good recognition rates [10], but speech quality degradation is a problem because the rules that are applyed are too coarse.

The paper is organized as follows: section 2 provides a brief state of the art of transplantation techniques and introduces the technique used by us in the cross-lingual transplantation, cross-lingual transplantation is introduced in section 3, and section 4 introduces the experimental framework applied to measure the problems of the proposed method. Finally in section 5 we give some brief conclusions and in section 6 we talk about the future work that we expect to carry out regarding the topic at hand.

2 Emotion Transplantation

Emotion transplantation methodologies can be defined as the procedures that allow the modification of a synthetic speech model to incorporate emotional information learned from other speaker models while maintaining the identity of the original speaker as much as possible. By this definition it follows that transplantation is a field of study that aims to solve one of the biggest problems in expressive speech synthesis: scalability.

A successful transplantation system should be capable of learning the paralinguistic nuances of the desired expressive speech model and then convert the target speaker model into a target speaker expressive model. But, as we would want the speaker identity to be maintained as much as possible, there are limitations on how much the features of the target speaker model should be modified [11]. In the end, systems have to reach a compromise between transplanted expressive strength and identifiability, synthetic speech quality and target speaker similarity. The considered transplantation technique attempts to combine the best features of both mainstream approaches to transplantation: adaptation to learn the function that converts a neutral model to substitute the rules while also being stable enough to provide good speech quality and complex enough to maintain the speaker identity.

2.1 Emotion Transplantation through Adaptation

The emotion transplantation process that we will apply in the proposed cross-lingual emotion transplantation system has been proven capable of correctly imbuing emotions into neutral speakers [5]. The system, whose flowchart can be seen in figure 1, relies on adaptation techniques for learning the transformation functions that characterize both speaker identity and emotional information. Then, the system applies both of them, and obtains the desired emotional target speaker model. All of this is done without requiring any kind of emotional information from the target speaker, which greatly helps reduce the scalability problems present in expressive speech synthesis.

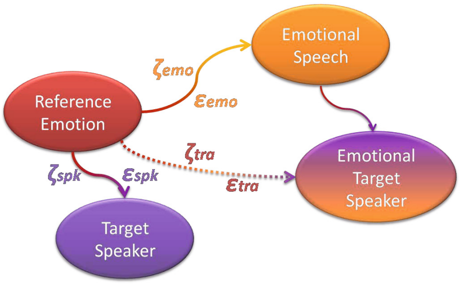

Fig. 1. Step by step block diagram of the emotion transplantation method. The spheres represent the speaker models and the arrows the adaptation transforms.

3 Cross-Lingual Emotion Transplantation

Cross-lingual processing is always difficult because of inter-language differences. Traditional approaches rely on phonetic mapping between the source and target

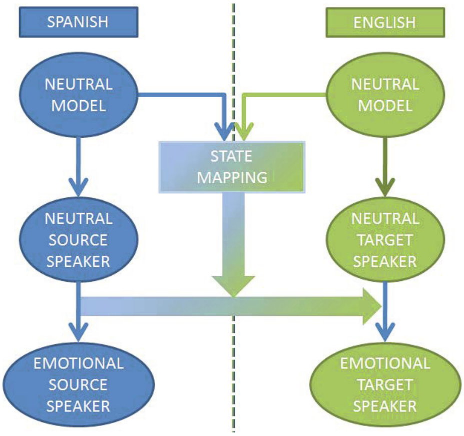

Fig. 2. Step by step block diagram of the emotion transplantation method. The spheres represent the speaker models and the thin arrows the adaptation transforms. The state mapping process is represented by the rectangle and the result of the mapping as the thick arrow. The two colors also represent the different languages.

languages at different levels in order to do a correspondence between what is done in the source language in hopes that if the same is done for the destination language, the results will be acceptable.

The first attempts at phonetic mapping relied on monophones or triphones without contextual information in order to establish direct relationships between languages [4,8], and evolved to using contextual information by exploiting the decision tree structures by measuring distances between tree nodes of both languages [14]. Different approaches tend to rely on having bilingual data for a single speaker and relying on state mapping and factor analysis techniques to extrapolate the knowledge that can be learned from the one speaker [16,17].

Typical applications of cross-lingual technologies include adapting speaker models between different dialects, accents, or variants of languages [12,13]. We believe that we can consider different expressiveness or speaking styles in the same fashion, and use our proposed transplantation technique combined with cross-language technologies with the purpose of transplanting paralinguistic features between languages.

3.1 Proposed Cross-Lingual Emotion Transplantation Method

The proposed cross-lingual emotion transplantation method is based on the state mapping principle [7]. The method begins by obtaining the mapping of the closest states amongst all states by means of the Karhunen-Loeve Divergence (KLD). Then, the emotional adaptation function obtained by means of the transplantation method is converted from the source model and source language to the target model and target language. Finally we apply the emotional transformation to the target language. A flowchart overviewing the complete process can be seen in figure 2.

4 Experimental Evaluation

Informal evaluations of the initial implementation of the cross-lingual emotion transplantation system showed that it is already capable of successfully transplanting the prosodic features between languages thanks to the state mapping technique, but a feature stream by feature stream transplantation showed that there is a significant amount of spectral noise introduced when transplanting the Cepstral information, specially in fundamentally prosodic emotions such as anger, which we will be using for the rest of the experimental evaluation section.

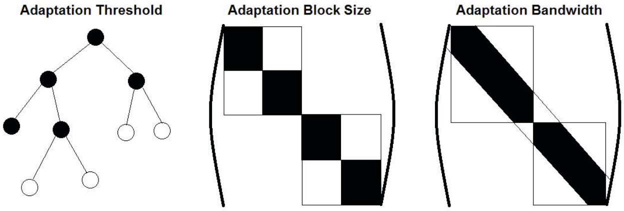

Fig. 3. Illustrations of the three different considered experimental variations. Black filled surfaces represent nonzero values.

In an attempt to quantify and lessen the effects of the introduced spectral noise we tried a number different experiments: increase the minimum number of frames per adaptation node, reduce the adaptation covariance transformation matrix block size and reduce the covariance adaptation bandwidth. We also developed an objective method to compare the results of the different approaches based on a combination of KLD and Principal Component Analysis (PCA). KLD is used to obtain the distances between all the speaker models used in the cross-lingual emotion transplantation process. PCA is used to project the KLD values into a bi-dimensional space that allows us to easily measure the distances in an objective way. This process is only done to the spectral features stream of the models as we want to measure the spectral distortion introduced to the process. In the ideal environment, the distance between the transplanted model (Korin-anger-S in the figures) and the target language model (Korin-neutral) should be the same as the distance between the source model (joa-anger-S) and the source language model (joa-neutral), which means obtaining a ratio of 1 in equation 1. The Spanish source data is a subsection of the SEV database [1], while the English target data was recorded in CSTR, University of Edinburgh.

4.1 Adaptation Function Occupancy Threshold

The first approach increased the adaptation function occupancy threshold. The purpose of this is to reduce the size of the decision tree used for the adaptation process in the source language by increasing the number of frames that must be present for each node (represented graphically in the first graph of figure 3). This produces more global transformation functions that convert the target language in a less complex fashion, potentially reducing the spectral noise introduced in the transplantation process. On the other hand, making use of more global adaptation functions means that there is less specificity in the transplantation, which could have an effect in the perceived emotion after transplanting.

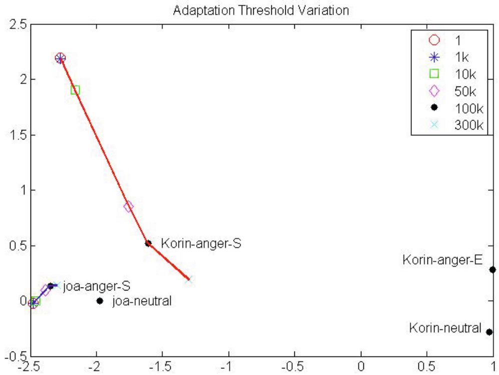

Fig. 4. PCA projection of the different adaptation threshold experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.

The results of this first experiments can be seen in figure 4, where it can be seen that while for lower threshold values there is no variation, for a very high value the distances between models clearly shrink, particularly between the emotional and neutral models of both languages. This means that using a higher threshold reduces the differences between the emotional and the neutral models.

4.2 Adaptation Variance Block Size

The second analysis aims to reduce the nonzero values in the adaptation function variance matrix by reducing the amount of intra-feature coefficients that influence each other (second illustration in figure 3). This means that, if in the standard block size every coefficient of a feature stream is able to influence each other, a smaller block size prevents the higher order coefficients from affecting lower ones.

The projection of the models obtained by manipulating the size of the blocks in the adaptation matrix can be seen in figure 5, and they show that both a

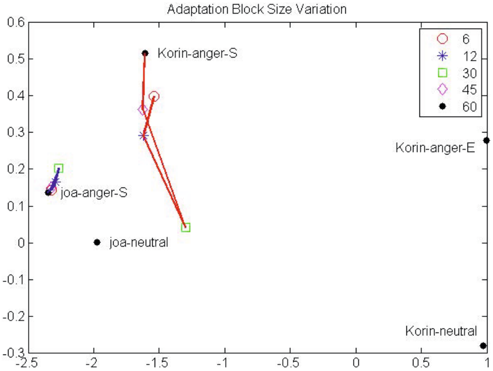

Fig. 5. PCA projection of the different adaptation block size experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.

big block size and a very small block size do not provide the optimal results, while an intermediate value provides less distant models. This is because a too small block size removes too many coefficients in the adaptation variance matrix, which results in a very coarse adaptation unable to deal correctly with emotional data [15].

4.3 Adaptation Variance Bandwidth

The final variant controls the values in the adaptation function variance matrix by only allowing a certain bandwidth of them around the diagonal to be nonzero as seen in the third illustration in figure 3. This approach also aims to reduce the interaction between higher and lower order coefficients in the adaptation of the feature streams, and it is supposedly smoother because there are no abrupt cuts in which coefficients affect each other. This approach makes special sense when using Linear Spectral Pairs (LSP) as the spectral features instead of Cepstral features, as the order of the coefficient correlates with the central frequency of the LSP. Thus, this approach directly limits the bandwidth that is considered in the adaptation process.

In the case of controlling the bandwidth of the adaptation variance matrix (figure 6) we can see that the lower the adaptation bandwidth the closest all the models become between each other. This translates into less noisy transplanted models but at the same time less expressive and identifiable models. This is also expected to give less expressive models as in the block size case.

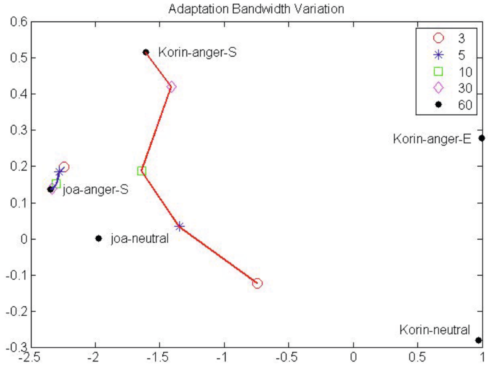

Fig. 6. PCA projection of the different adaptation bandwidth experiments in the anger cross-lingual extrapolation between joa (Spanish speaker) and Korin (English speaker). The prefix S stands for Spanish and E for English source data. The black dot is the reference system constant across experiments.

4.4 Perceptual Evaluation and Covariance Analysis

In an attempt to verify the validity of the objective measures, we carried out a small perceptual evaluation which aimed to obtain subjective speech quality results for all the considered systems. In this evaluation we presented the listeners with 14 different audio samples (one per system and only one at a time), and asked them to rate the perceived speech quality from 1 (very bad) to 5 (very high). A total of 14 utterances were synthesized according to the Latin square evaluation strategy in order to remove any bias in the evaluation process. Finally, we measured the correlation between a ratio of the model distances and the speech quality as follows:

$$corr(\frac{dist(TransplantedEnglish_i - NeutralEnglish)}{dist(EmotionalSpanish_i - NeutralSpanish)}, SpeechQuality_i) \qquad (1)$$

In the results of measuring said correlation (table 1) we can see that there is a very strong relationship between the ratio of the distances between transplanted and source models and the perceived speech quality. This means that the proposed objective measure is a good tool of measuring how much we can improve speech quality compared to the initial cross-lingual transplantation system.

Table 1. Covariance analysis results for the three experimental environments

Evaluation System	Threshold	Block Size	Bandwidth
Measured Correlation	-0.948	-0.952	-0.893

5 Conclusions

We have set the foundations for a cross-lingual emotion transplantation system, where we are able to successfully convert the adaptation functions that convey emotional information from a source language (Spanish) to the desired target language (English) by means of a state mapping technique based on minimizing the KLD between neutral language models. This has enabled us to correctly modify the prosody and spectral components of the target speaker in the target language in the same fashion that we would have done in the source language, conveying then the desired emotion.

Preliminary evaluations show that there is a significant amount of spectral distortion introduced in the transplantation process. By measuring the KLD distance and projecting it into a bi-dimensional space with PCA we plotted the distances between the source and the transplanted models, providing us with an easy way of measuring the effects of our experiments. The experiments aimed to reduce the spectral noise by means of controlling different adaptation parameters such as the adaptation threshold or limiting the non-zero values of the adaptation variance matrix. The experiments showed that we should minimize the distances across neutral speaker models in both languages in order to obtain a really successful emotion transplantation across languages. We have also carried out a perceptual evaluation of the speech quality of the transplanted models that, by means of a correlation analisis, helped prove that the proposed objective measure is a good way of visualizing the transplantation results.

6 Future Work

Future work can be separated into two sections: first of all we want to record a bilingual neutral corpus in a controlled environment that we can use to evaluate the successfulness of the proposed technique when there are no environmental differences present. Secondly we also want to try different ways of minimizing the model variability such as applying Cepstral variance and mean normalization to the neutral speakers, or implementing environmental factoring in the average voice model training process.

Acknowledgements. Part of this work has been funded by, and carried out during the research stay of Jaime Lorenzo-Trueba at the National Institute of Informatics (NII) in Spring 2014. The work leading to these results has also received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02), and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo has been funded by Universidad Politécnica de Madrid under grant SBUPM-QTKTZHB. The authors want to thank the other members of the Speech Technology Group, ARABOT and Simple4All projects for the continuous and fruitful discussion on these topics.

References

1. Barra-Chicote, R., Montero, J.M., Macias-Guarasa, J., Lufti, S., Lucas, J.M., Fernandez, F., D'haro, L.F., San-Segundo, R., Ferreiros, J., Cordoba, R., Pardo, J.M.: Spanish expressive voices: Corpus for emotion research in spanish. In: Proc. of LREC (2008)
2. Barra-Chicote, R.: Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis. Ph.D. thesis, ETSIT-UPM (2011)
3. Gales, M.J.: Cluster adaptive training of hidden markov models. IEEE Transactions on Speech and Audio Processing 8(4), 417–428 (2000)
4. Liang, H., Dines, J.: Phonological knowledge guided hmm state mapping for cross-lingual speaker adaptation. In: INTERSPEECH, pp. 1825–1828 (2011)
5. Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., Montero, J.M.: Towards speaking style transplantation in speech synthesis. In: 8th ISCA Speech Synthesis Workshop (2013)
6. Nose, T., Kato, Y., Kobayashi, T.: Style estimation of speech based on multiple regression hidden semi-markov model. In: INTERSPEECH, pp. 2285–2288 (2007)
7. Oura, K., Yamagishi, J., Wester, M., King, S., Tokuda, K.: Analysis of unsupervised cross-lingual speaker adaptation for hmm-based speech synthesis using kld-based transform mapping. Speech Communication 54(6), 703–714 (2012)
8. Qian, Y., Xu, J., Soong, F.K.: A frame mapping based hmm approach to cross-lingual voice transformation. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5120–5123. IEEE (2011)
9. Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Eigenvoices for hmm-based speech synthesis. In: INTERSPEECH (2002)
10. Takeda, S., Kabuta, Y., Inoue, T., Hatoko, M.: Proposal of a japanese-speech-synthesis method with dimensional representation of emotions based on prosody as well as voice-quality conversion. International Journal of Affective Engineering 12(2), 79–88 (2013)
11. Togneri, R., Pullella, D.: An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine 11(2), 23–61 (2011)
12. Toman, M., Pucher, M., Schabus, D.: Multi-variety adaptive acoustic modeling in hsmm-based speech synthesis. In: 8th ISCA Speech Synthesis Workshop (2013)
13. Toman, M.E., Pucher, M.: Structural kld for cross-variety speaker adaptation in hmm-based speech synthesis. In: Proc. SPPRA, Innsbruck, Austria (2013)
14. Wu, Y.J., Nankaku, Y., Tokuda, K.: State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis. In: INTERSPEECH, pp. 528–531 (2009)
15. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. IEEE Transactions on Audio, Speech, and Language Processing 17(1), 66–83 (2009)
16. Yoshimura, T., Hashimoto, K., Oura, K., Nankaku, Y., Tokuda, K.: Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for hmm-based speech synthesis. In: 8th ISCA Speech Synthesis Workshop (2013)
17. Zen, H., Braunschweiler, N., Buchholz, S., Knill, K., Krstulovic, S., Latorre, J.: Hmm-based polyglot speech synthesis by speaker and language adaptive training. In: Seventh ISCA Workshop on Speech Synthesis (2010)