Towards Speaking Style Transplantation in Speech Synthesis

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

[1]Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain
[2]CSTR, University of Edinburgh, United Kingdom
{jaime.lorenzo, barra}@die.upm.es

Abstract

One of the biggest challenges in speech synthesis is the production of naturally sounding synthetic voices. This means that the resulting voice must be not only of high enough quality but also that it must be able to capture the natural expressiveness imbued in human speech. This paper focus on solving the expressiveness problem by proposing a set of different techniques that could be used for extrapolating the expressiveness of proven high quality speaking style models into neutral speakers in HMM-based synthesis. As an additional advantage, the proposed techniques are based on adaptation approaches, which means that they can be used with little training data (around 15 minutes of training data are used in each style for this paper). For the final implementation, a set of 4 speaking styles were considered: news broadcasts, live sports commentary, interviews and parliamentary speech. Finally, the implementation of the 5 techniques were tested through a perceptual evaluation that proves that the deviations between neutral and speaking style average models can be learned and used to imbue expressiveness into target neutral speakers as intended.

Index Terms: expressive speech synthesis, speaking styles, adaptation, expressiveness transplantation

1. Introduction

Speech synthesis is a field that has been seeing much more use in the last decade with the advent of human-machine interfaces, playing an integral role in them. As such there have been constant studies on how to improve its quality, naturalness, expressiveness, etc. Among these efforts is the project under which this investigation is enclosed: Simple4All [1]. Simple4All is an European funded project whose main purpose is to streamline the training process of expressive synthetic voices by creating a system that requires little to no supervision and is capable of learning constantly just by its interactions with the users.

More concretely, expressive speech synthesis is a sub-field of speech synthesis that has been drawing a lot of attention lately, as until recently there was no effort paid to increasing the adequacy of the produced voices to the task they were intended to be used in. But, if one were to assign expressiveness to the synthetic voices (e.g. emotions or speaking styles), the result would be a much more natural voice increasing the overall satisfaction of the end users of the interface. This, when considering the two main speech synthesis techniques (unit selection and HMM-based) places a serious restriction that clearly favors HMM-based synthesis [2]: if expressive data were to be recorded for every possible situation, the size of the databases would become immense, making unit-selection nonviable on principle. HMM-based synthesis, on the other hand, due to its parametric nature is much more adaptable, a fact that can be exploited even further by using adaptation techniques [3].

Consequently this study focuses on HMM-based synthesis and adaptation techniques in order to produce voices with the desired speaking styles. Firstly, and keeping in mind that the final training system should require minimal interaction from the user, it is interesting to minimize the training data required to produce the output models without reducing the final quality. This can be done by exploiting background average models [4] from which the final voice is adapted using one of the different available techniques (this study relies on CSMAPLR adaptation [3]). At this point the problem becomes how to imbue the models with speaking styles, towards which we can see some recent studies such as Cluster Adaptive Training [5], that relies on clustering the expressive training speakers into a continuous expressive speech space of the different available speaking styles.

The approach suggested in this paper consists of creating representative models for every desired expression from small subsets of data that clearly show the nuances of that particular oratory, including one for neutral or read-speech voices. Then we propose a way of modeling the differences between the neutral model and the target speaking style through adaptation transformations, together with a way of transferring this differences into a new neutral target speaker in order to adapt the voice into the desired speaking style pattern. This in the end allows us to generate voices with speaking styles for any target neutral speaker even if there is no previous expressive data available for that speaker. This technique is finally verified through a perceptual test, showing the usefulness of extrapolating the speaking style of average models as the results are considered by the listener to be significantly more adequate to the proposed speaking styles than the traditional neutral voices.

2. Speech Corpora and Average Models

For both the speaking styles and neutral read speech corpora we used a combination of pre-existent databases and some new recorded hand-labeled data, separated as follows:

2.1. Speaking Styles Speech Data

C-ORAL-Rom Database A multi-language multi-style database [6]. Out of all the available data, only three of the styles available in the Spanish formal media section were used: news, sports and interviews, and between all the available data of each style a subset of the least noisy audio files was selected.

TC-STAR run 3 A multi-language database of recorded parliamentary speeches in different environments [7] such as the European Parliament or the Spanish Parliament. Out of all the available data, four different speakers in the Spanish Parliament subsection were used.

Self Labeled Data Because some of the styles did not amount to enough data (namely: news and sports), additional speech was processed and added to the models. For the news style, recorded data of live news by a very famous Spanish newscaster was processed. Finally, for the sports commentary style, we aligned and labeled 15 minutes of the broadcast of the Eurocup2012 finals.

2.2. Neutral Read Speech Data

UVIGO-ESDA Database A database consisting of a single male amateur Spanish speaker (UVD) in a neutral read speech situation for approximately 2 hours of speech recorded in studio [8]. This speaker was used for obtaining both the average modeling and also for the implementations of the speaking styles extrapolation techniques.

SEV Database An emotional database consisting of a male and a female speaker [9]. Only the neutral speech of the male speaker was used, and only for the average modeling.

New Recorded Data In order to increase the variability of the neutral read speech data we also added a few speakers of those previously recorded in our laboratory environment. The recording is done inside an acoustically-treated room, so the obtained quality is very high. Out of all this data, 4 male speakers were added to the average model data pool and 2 of them, one possessing a mid-range pitch (JLC) and a final one with a high-range pitch and a soft Colombian accent (JEC) were used for the final synthesis and perceptual test.

With all the mentioned data and by using Speaker Adaptive Training (SAT) [4] a complete background average model was obtained. This model, which contains both neutral and speaking styles data, will be used as the basis of all further adaptations, significantly reducing the final voices training time and increasing the overall quality and robustness of the models as proved in some of our previous work [10].

2.3. Adapted Average Models

As the full background model is too complex to be able to capture particular nuances of the different expressive styles, we applied an intermediate adaptation step with which we obtained average models of the 4 speaking styles (sports, news, interviews, politics) and an additional one for the neutral read speech speakers. This adaptation was done by using the CSMAPLR algorithm [3], chosen because of its synergy with SAT models and high quality adaptation even with the small adaptation data available for the speaking styles models. The biggest advantage of having average models of the 5 speaking styles considered is that the differences between the pure styles can be characterized and exploited for the expressiveness transplantation, as will be exploited and explained later in this paper.

Finally, the speaker models of the neutral speakers used for the perceptual test were adapted using CSMAPLR again but in this case directly adapting from the neutral average, which will be an important detail when facing the speaking styles adaptation through transplantation that will be defined in the next section.

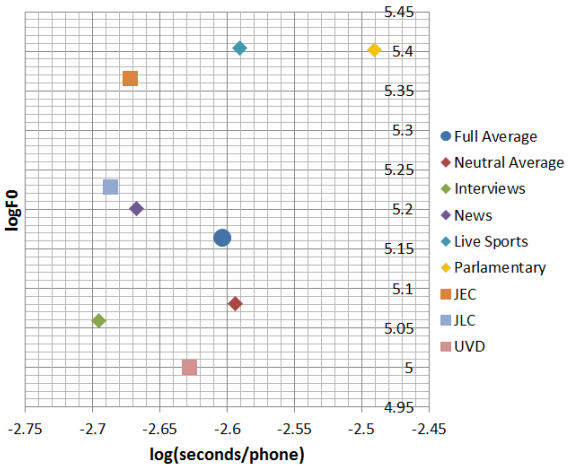

Figure 1: Prosody characterization of the training dataset. Diamonds represent the average models and squares the neutral speakers. Uttering speed was selected to obtain the faster models to the right while keeping them separable.

2.4. Analysis of the Training Data

In figure 1 we can see a plot of two main prosody features (logF0 and -1/uttering speed) of the data used to train the average models. In it we can define 3 F0 bands around the 3 test speakers, a low-range F0 for UVD, mid-range F0 for JLC and a high band for JEC. Similarly we can consequently associate also each style to a band: interviews to low-F0, news to mid-range F0 and live sports and parliamentary speech to high F0. This was already observed in a previous study [11]. That is, a live sports commentary or a parliamentary speech will be typically recorded in an open, noisy environment while an interview or a newscast will be recorded on studio. Uttering speed can be explained by the spontaneity of the style: a newscaster will have a prepared script that can be read quickly while a politic will tend to somewhat improvise on the reactions of the public.

3. Speaking Styles Transplantation

The main objective of the present research is to be able to transplant the nuances of particular speaking styles speech models into a different neutral speaker. This would imply a significant step up in the availability of task-dependent voices, which is much needed when considering naturalness of the synthetic speech. Additionally, because the proposed techniques are based on adaptation the harvesting of data becomes much simpler, as it is possible just to pick speaking styles examples of a particular style or emotion of different speakers from any source in order to train the average models used as the basis for the extrapolation. This in the end means a substantial increase in the ease of producing synthetic voices with speaking styles.

3.1. Proposed Approach: Transplantation through Adaptation

One of the biggest advantages of parametric speech synthesis is its versatility, and adaptation is a technique that exploits that versatility successfully. It can then be used to obtain robust models from a background average and a few minutes of the speaker, and also to obtain the transformation functions between models to help characterize the differences between them.

This principle was applied in our Albayzin2012 speech synthesis challenge submission [10] to successfully control the expressive strength of emotional models by assuming that the transformation relating the expressive and neutral model can be scaled, allowing for a linear continuous modeling of the expressiveness space.

Following up on the strength control concept the concept of expressiveness extrapolation appears. If the transformation function between an expressive model and a reference model can be transferred to a different speaker, it is natural to think that the expressiveness will be likewise transferred to the target speaker.

It is not acceptable to think that the relationship between a particular speaker's expressive representation and that same speaker's neutral voice is the real representation of that expressiveness, and that is why we propose the use of averages. If we can model the transformation between the target speaker and the neutral average and apply this transformation to the expressive average, it is to be expected that we will obtain the desired target expressive speaker (figure 2).

Both the adaptation process between the background averages and the adaptation between the neutral average and the neutral target speaker are done using CSMAPLR. This presents the advantage of providing linear transformations that consider both mean and variances. If the adaptation were not to be constrained the variances would be ignored and the expressive nuances could be lost [3], and the linearity of the transformation reduces the complexity of the modeling.

Both CSMAPLR adaptation transformations can be expressed as :

$$\bar{\mu}_{exp} = \zeta_{exp}\mu_N + \epsilon_{exp} \quad (1)$$

$$\bar{\Sigma}_{exp} = \zeta_{exp}\Sigma_N\zeta_{exp}^T \quad (2)$$

$$\bar{\mu}_{spk} = \zeta_{spk}\mu_N + \epsilon_{spk} \quad (3)$$

$$\bar{\Sigma}_{spk} = \zeta_{spk}\Sigma_N\zeta_{spk}^T \quad (4)$$

Where $\bar{\mu}_{exp/spk}$ and $\bar{\Sigma}_{exp/spk}$ are the target means and covariance matrices of the expressive and target speaker models respectively, with ζ defining the rotation matrix and ϵ the bias that are obtained following the CSMAPLR algorithm [3]. Consequently, the transplantation transform is defined as follows:

$$\bar{\mu}_{tra} = \zeta_{spk}\zeta_{exp}\mu_N + \zeta_{spk}\epsilon_{exp} + \epsilon_{spk} \quad (5)$$

$$\bar{\Sigma}_{tra} = \zeta_{spk}\zeta_{exp}\Sigma_N\zeta_{exp}^T\zeta_{spk}^T \quad (6)$$

3.2. Alternatives to Transplantation: Copying the Speaking Style Average Model

In order to test the relevance of the proposed transplantation adaptation technique we defined a set of alternatives that could be considered for expressive synthesis, namely copying the different feature streams from the average models into the target speaker model. In this case it meant copying either the prosody features (F0 and duration streams) or the spectral features. This would not be an easy thing to do in a situation in which every voice was trained independently following the traditional HMM-based modeling, as the decision trees would not be shared. But because our voices share a common background model and the adaptation process keeps the trees intact, copying the prosody or the spectrum from the style averages is as simple as replacing the desired model files in the target speaker.

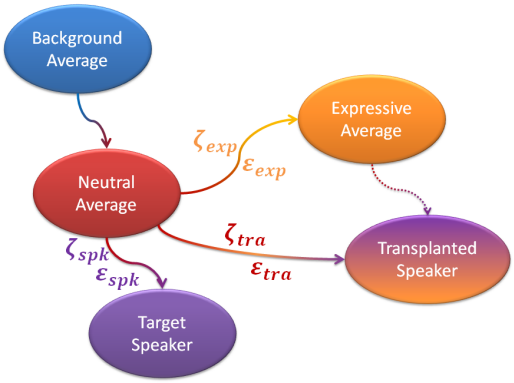

Figure 2: Schematic of the extrapolation through adaptation system.

3.2.1. *Copying the Prosody*

Prosody is known to carry a very significant portion of the expressive load of speech [12]. As such it is assumable that by simply copying the prosody (only F0 and phone durations are considered) of a clear representative of a speaking style model would yield acceptable extrapolation results at least if the target speakers' F0 and the average's F0 are not too dissimilar. On the other hand, it is also possible that if this last condition does not hold the quality of the output voice would degrade or that instabilities might appear.

3.2.2. *Copying the Spectrum*

Spectrum is assumed to hold most of features related to the identifiability of the speaker, although it also includes some expressive features [12]. In that sense it is safe to assume that this approach would not be very successful for extrapolating the expressiveness of the model but instead copy the identity of the style's average. This by itself does not seem useful for the task at hand, although for example in some extreme cases were a speaker can be clearly associated to a particular style it would fulfill a similar purpose. Nonetheless, it seems interesting to test the results this kind of approach would give.

4. Perceptual Test Description

To test the effectiveness of the proposed technique we prepared a web-based perceptual test in which 32 listeners were asked to establish two different rankings in order of preference: adequacy of the speaking style to the synthesized text and similarity to the original speaker. The decision to make the test ranking-based was taken because, as the task is considerably difficult and there is no natural voice reference available for the listener, comparing the system between themselves instead of assigning a value to them facilitated the testing process.

The target's speaker neutral voice and the speaking style average model were synthesized to be added as top-line systems: the neutral voice would provide the top-line for the similarity analysis and the average model for the adequacy of the speaking style task.

The test consists of 5 systems (top-line systems, transplantation system, copy-prosody system and copy-spectrum system) and 4 styles (news, sports, interviews and parliamentary speech) for each of the 3 evaluated target speakers.

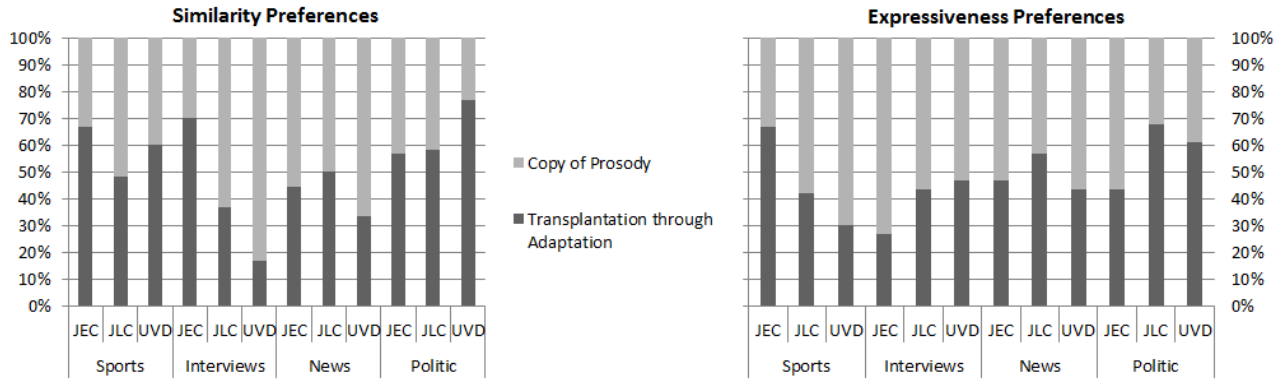

Figure 3: Results of the perceptual test comparing the user preferences between transplantation and copy of prosody for all test speakers and styles.

Regarding the interface, the listeners were presented with all five stimuli at the same time so that they could be played as many times as necessary. The stimuli distribution was designed following a balanced latin square randomization of the questions pattern, resulting in a total of 8 utterances required per style. The utterances were extracted from real media texts and were selected so as to be longer than 10 words for the phrase-level prosody to become relevant. Finally, the minimum acceptable number of tests per target speaker was decided to be 16: two complete rounds of tests.

5. Results

An initial consideration verified by table 3 is that copying the spectrum of the average or using the average itself is not a valid transplantation technique because these methods do not keep the identity of the target speaker, even if the perceived adequacy is comparable between the 5 proposed systems (table 2). As such the analysis will focus only on when is adaptation considered by the listeners better than copy of prosody and vice versa.

The first result that becomes evident from both table 1 and figure 3 is that there is no significant difference overall between transplantation and extrapolating by copying the prosody of the average model. Even so, in a global level it can be seen that specially the parliamentary speech style but also the live sports commentary (the higher F0 ones, as seen in figure 1) favor the transplantation-based system while news favors the copy of prosody in the similarity department, without relevant differences in adequacy. This is reinforced by the similarity preference results of JEC, the high pitched target speaker, for which the results definitely show that transplantation through adaptation is the preferred technique.

The consideration to be drawn from the adequacy results is that even if in average the test results do not favor any of the techniques, the more different speaker-style pair of models (i.e. UVD with politics or JLC with politics) also appear to be more adequate for transplantation than for copy of prosody. This could be seen as a hint that while copying the prosody is an acceptable method of extrapolating the speaking style, it stops being reliable when the pair of models is too disparate in prosody. On the other hand, transplantation through adaptation does not fall off in these kind of situations because not only the prosody is adapted but also the spectrum of the models, preventing instabilities or unnatural sounds from appearing.

Table 1: Number of utterances preferred (>) by listeners in terms of expressiveness and similarity between transplantation-based (Trans) and copy prosody-based (C-Pro) systems.

Expressiveness	SIMILARITY		TOTAL
	Transp<C-Pro	Transp>C-Pro	
Transp<C-Pro	**98**	**91**	**189**
Sports	22	27	49
Interviews	30	25	55
News	29	17	46
Politic	17	22	39
Transp>C-Pro	**76**	**98**	**174**
Sports	16	26	42
Interviews	23	12	35
News	21	23	44
Politic	16	37	53
TOTAL	**174**	**189**	**363**

6. Conclusions

The first conclusion that can be drawn from the test is that the extrapolation of speaking styles can provide synthetic voices more adequate to different tasks (i.e. style of delivery) than simpler neutral voices for any speaker without requiring the target speaker to record any non-neutral data. This is a huge step-up from the traditional synthesis algorithms that would require the target speaker to record a new database for every expressive realm they want their voice on.

Also, we have seen that using average models allows this extrapolation to be done with as little as 15 minutes of speaking styles data. It has been done both by copying these average models' prosody or by applying the more advanced technique of adapting between neutral and speech with speaking styles to the neutral speaker. In general both techniques appear to be capable of imbuing the target voices with speaking styles while keeping the source identity, but we have found a trend in which target voices that are too different from the average models start producing worse quality voices when just copying prosody.

Additionally, when considering different applications such as emotional speech synthesis, merely copying the prosody will not be able to extrapolate the expressiveness in all situations, requiring a more complex approach such as our transplantation process.

Table 2: Results in adequacy ranking for the different systems averaged between the 3 target speakers.

ADEQUACY	Read Speech	Style Average	Copy of Spectrum	Copy of Prosody	Transplantation
Sports	2.57	2.50	2.15	3.96	3.81
Interviews	2.63	3.69	2.98	2.94	2.76
News	2.98	3.22	2.68	3.37	2.76
Politic	3.49	2.82	2.40	3.02	3.26
Average	2.92	3.06	2.55	3.32	3.15

Table 3: Results in similarity ranking for the different systems averaged between the 3 target speakers.

SIMILARITY	Read Speech	Style Average	Copy of Spectrum	Copy of Prosody	Transplantation
Sports	4.31	1.85	2.36	3.13	3.35
Interviews	4.50	1.68	2.28	3.37	3.18
News	4.36	1.89	2.12	3.41	3.22
Politic	4.40	1.73	2.35	3.01	3.51
Average	4.39	1.79	2.28	3.23	3.32

Even so, the results are not significant enough yet, so the planned future work is two-fold: first of all increase the available training data for each speaking style so as to obtain much more informative averages from which to adapt. We also plan to add strength control capabilities to the transplantation adaption system in order to try different control ratios to try and find optimal control values that increase the perceived adequacy of the style by enhancing particular features that carry more expressiveness information. Finally, we intend to test the proposed system in an emotional environment and compare it with the considered systems once again to verify the versatility we can provide.

7. Acknowledgments

The work leading to these results has 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 Politecnica de Madrid under grant SBUPM-QTKTZHB. Authors also thank the other members of the Speech Technology Group and Simple4All project for the continuous and fruitful discussion on these topics.

8. References

[1] Rob Clark and Simon King, "Simple4all - http://simple4all.org," 2011.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[3] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[4] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[5] Langzhou Chen, Mark Gales, Vincent Wan, Javier Latorre, and Masami Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13*, 2012.

[6] A. Moreno-Sandoval, G. De la Madrid, M. Alcántara, A. Gonzalez, JM Guirao, and R. De la Torre, "The spanish corpus," *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, Amsterdam: John Benjamins Publishing Company*, pp. 135–161, 2005.

[7] A. Cardenal-Lopez L. Docio-Fernandez and C. Garcia-Mateo, "Tc-star 2006 automatic speech recognition evaluation: The uvigo system," pp. 145–150, 2006.

[8] C.G. Mateo E.T. Banga, "Documentation of the uvigo-esda spanish database," Tech. Rep., Grupo de Tecnoloxias Multimedia, Universidad de Vigo, Vigo, Espaa, 2010.

[9] R. Barra-Chicote, J. M. Montero, J. Macias-Guarasa, S. Lufti, J. M. Lucas, F. Fernandez, L. F. D'haro, R. San-Segundo, J. Ferreiros, R. Cordoba, and J. M. Pardo, "Spanish expressive voices: Corpus for emotion research in spanish," *Proc. of LREC*, 2008.

[10] Jaime Lorenzo-Trueba, Oliver Watts, Roberto Barra-Chicote, Junichi Yamagishi, Simon King, and Juan M. Montero, "Simple4all proposals for the albayzin evaluations in speech synthesis," in *Iberspeech2012, VII Jornadas en Tecnologia del Habla and III Iberian SLTech Workshop*, 2012.

[11] Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Tuomo Raitio, Nicolas Obin, Paavo Alku, Junichi Yamagishi, and Juan M Montero, "Towards glottal source controllability in expressive speech synthesis," in *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13*, 2012.

[12] Roberto Barra-Chicote, *Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis*, Ph.D. thesis, ETSIT-UPM, 2011.