Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 292–307

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Emotion transplantation through adaptation in HMM-based speech synthesis[☆]

Jaime Lorenzo-Trueba[a,*], Roberto Barra-Chicote[a], Rubén San-Segundo[a], Javier Ferreiros[a], Junichi Yamagishi[b], Juan M. Montero[a]

[a] *Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Avenida Complutense n° 30, Ciudad Universitaria, 28040 Madrid, Spain*
[b] *CSTR, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom*

Received 24 May 2014; received in revised form 3 February 2015; accepted 25 March 2015
Available online 3 April 2015

Abstract

This paper proposes an emotion transplantation method capable of modifying a synthetic speech model through the use of CSMAPLR adaptation in order to incorporate emotional information learned from a different speaker model while maintaining the identity of the original speaker as much as possible. The proposed method relies on learning both emotional and speaker identity information by means of their adaptation function from an average voice model, and combining them into a single cascade transform capable of imbuing the desired emotion into the target speaker. This method is then applied to the task of transplanting four emotions (anger, happiness, sadness and surprise) into 3 male speakers and 3 female speakers and evaluated in a number of perceptual tests. The results of the evaluations show how the perceived naturalness for emotional text significantly favors the use of the proposed transplanted emotional speech synthesis when compared to traditional neutral speech synthesis, evidenced by a big increase in the perceived emotional strength of the synthesized utterances at a slight cost in speech quality. A final evaluation with a robotic laboratory assistant application shows how by using emotional speech we can significantly increase the students' satisfaction with the dialog system, proving how the proposed emotion transplantation system provides benefits in real applications.

Keywords: Statistical parametric speech synthesis; Expressive speech synthesis; Cascade adaptation; Emotion transplantation

1. Introduction

ARABOT and INAPRA (and previously URBANO and ROBONAUTA) are coordinated Spanish research projects on interactive mobile robotics for real environments such as museums or universities. The robots integrate autonomous navigation, a distributed object-oriented architecture, automatic speech recognition, affective speech synthesis, a mechatronic emotional face and robotic arms (Rodriguez-Losada et al., 2008). In order to adapt to an ever-changing domain

[☆] This paper has been recommended for acceptance by Roger K. Moore.

[*] Corresponding author. Tel.: +34 91 5495700.

E-mail addresses: jaime.lorenzo@die.upm.es (J. Lorenzo-Trueba), barra@die.upm.es (R. Barra-Chicote), lapiz@die.upm.es (R. San-Segundo), jfl@die.upm.es (J. Ferreiros), yjamagis@inf.ed.ac.uk (J. Yamagishi), juancho@die.upm.es (J.M. Montero).

http://dx.doi.org/10.1016/j.csl.2015.03.008

of application, the robot has a domain-independent emotional model of behavior (Lutfi et al., 2013) which is able to automatically estimate the degree of satisfaction of the users the robot is interacting with, and is able to adapt the emotional state and the spoken dialog to the context of use. By means of this adaptive empathetic strategy, the artificial agent significantly increases users' satisfaction and minimizes users' frustration, even when the performance of the speech recognizer or the dialog manager cannot be improved.

Current speech synthesis systems, whether we are talking about unit selection or HMM-based systems, can provide very good naturalness and intelligibility when synthesizing read speech regardless of the technology (Barra-Chicote et al., 2010; Barra-Chicote, 2011) which is ideal for neutral speech interfaces that do not need to engage in a direct conversation with the user. On the other hand, applications such as dialog systems (Lutfi et al., 2013), robots or virtual characters, where simulating a more human-like behavior is necessary, a neutral speech synthesis does not live up to the task. Imbuing the synthetic speech with expressive features (e.g. emotions, speaking styles, etc.) is the role of expressive speech synthesis.

Due to the sheer amount of possible expressiveness, recording complete databases that cover all of them is unthinkable, making unit selection based systems fall behind in terms of scalability, although they are definitely capable of producing expressive speech (Adell et al., 2010, 2012; Andersson et al., 2010; Erro et al., 2010). On the other hand, HMM-based systems, because of their parametric nature, can be easily adapted through speaker adaptation techniques and can be successfully used for this task, and have been proven to provide significant improvements in perceived speech quality (Yamagishi et al., 2005).

One of the biggest problems of expressive speech synthesis is data acquisition. As human expressiveness is not a discrete space but a continuous one, the expressive strength and nuances vary greatly not only from person to person but from utterance to utterance for the same person. This problem can be focused on from different approaches: lexical analysis (Andersson et al., 2012) for correctly classifying the available data and training more precise systems or acoustic analysis. For acoustic analysis several aspects have been considered such as expressiveness detection (El Ayadi et al., 2011; Lorenzo-Trueba et al., 2012; Schuller et al., 2010), expressiveness production (Obin et al., 2011; Raitio et al., 2013), expressive intensity control (Nose et al., 2013; Picart et al., 2011) or expressiveness transplantation (Chen et al., 2012; Latorre et al., 2012).

The work present in this paper is enclosed mainly under the field of expressive speech synthesis, and aims to fix one of its main shortcomings: scalability. Human communication is so rich and so deep that it is impossible to imagine obtaining data for every combination of speaker and expressiveness, and that is why we want to propose a method capable of learning the paralinguistic information of emotional speech, control its emotional strength and transplant it to different speakers for whom we do not have any expressive information. We decided to focus on emotional speech as a particularization of expressive speech (fitting the aim of creating different emotional voices for several affective robots in a museum such as the Principe Felipe Science Museum in Valencia we are collaborating with), but we can expect the transplantation method to be able to support different expressive domains.

A successful transplantation method that has been introduced lately (Chen et al., 2012; Latorre et al., 2012) is based on Cluster Adaptive Training (CAT) (Gales, 2000), a projective adaptation technique. As such it is only capable of producing speaker models based on linear combinations of the original training speaker models. The main advantage of this approach is that as the produced model is always a combination of pre-existing training models, the process is extremely robust, outputting very high quality speech (Yanagisawa et al., 2013). On the other hand, the level of expressive strength or speaker similarity cannot be guaranteed as the transplantation reach is very constrained. This is also the case for model interpolation techniques (Hsu et al., 2012), capable of achieving better expressiveness than traditional adaptation techniques at a cost in speaker similarity.

Another approach to emotion transplantation is the use of rules to directly modify the synthesis models. This approach is theoretically capable of imbuing an emotion on any target speaker as long as we know the correct rules. In reality these approaches, while usually capable of providing emotional strength controllability and reasonably good recognition rates (Zovato et al., 2004; Takeda et al., 2013), speech quality and speaker similarity degradation tend to be a problem.

The proposed emotion transplantation method considers the best of both previously mentioned approaches: using adaptation to lessen speech quality degradation while using the adaptation functions as pseudo-rules for modifying the speaker models. As a result we present a method capable of controlling expressive strength while reasonably maintaining speech quality and speaker identifiability when compared to non-transplanted expressive synthetic speech (Lorenzo-Trueba et al., 2013a,b).

Table 1
Complete characterization of the speakers used in the evaluations. In the emotions section, A stands for anger, H for happiness, S for sadness, U for surprise and N for neutral.

Speaker	Gender	Emotions	Amount of data
JOA	Male	A, H, S, U, N	2 h 30 min
ROS	Female	A, H, S, U, N	2 h 30 min
UVD	Male	N	2 h
JEC	Male	N	7 min
JLC	Male	N	3 min
UEM	Female	N	1 h 45 min
NAS	Female	N	7 min
EMA	Female	N	6 min

The paper is organized as follows. In Section 2 we introduce the neutral and emotional speech corpora we have used for training and evaluation purposes during the development of the proposed method. Section 3 introduces the transplantation method, where Section 3.1 introduces the mathematical aspects of the used CSMAPLR adaptation and how it was expanded for our purposes, and Section 3.2 explains in detail the procedure through which the emotion transplantation is applied, together with an emotion transplantation baseline. Section 4 describes how the perceptual evaluations were carried out and analyzes the results. In Section 5 we describe and present the results of our evaluation with a robotic agent. Finally in Section 6 we discuss the results obtained in our evaluations and in Section 7 we present the conclusions to be drawn from this paper together with a brief summary of the main proposals.

2. Speech corpora

For the development and evaluation of the proposed emotional speech transplantation method we employed both neutral and emotional databases. The emotional database (SEV, Barra-Chicote et al., 2008) was evaluated in the Albayzin 2012 speech synthesis challenge. The neutral data is a combination of databases from previous Albayzin challenges (Mendez Pazo et al., 2010) (UVIGO-ESDA Database, Banga, 2010 and UPC-ESMA Database, Bonafonte and Moreno, 2008) and a number of male and female speakers recorded in our laboratory environment.

SEV Database Emotional database consisting of a male and female speaker. Out of the available emotions only 4 of them were considered: anger, happiness, sadness and surprise also including the neutral voice as the reference. All the emotions were recorded for the same utterances favoring the learning of expressiveness cues. There is approximately 30 min of training speech for each emotion and speaker.

UVIGO-ESDA Database A database consisting of a single male Spanish speaker (UVD) in a neutral situation totaling 2 h of speech recorded in studio.

UPC-ESMA Corpus A database consisting of a single professional female speaker (UEM) totaling 1.75 h of neutral style speech, recorded in a noise reduced room.

Recorded Data A number of male and female speakers were recorded in our acoustically-treated room, providing high quality and stable speech. Two male (JLC and JEC) and two female speakers (NAS and EMA) were used as the transplantation targets. Data durations vary from 6 min for EMA, 7 min for JEC to 30 min for JLC.

The corpora can be seen in Table 1, where we can see the acronyms that identify the speakers, their gender, emotions and amount of data.

3. 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. 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.

3.1. Adaptation-based transplantation

Adaptation is a powerful tool when considering emotional speech synthesis and more concretely emotion transplantation, as it allows us to exploit the versatility of HMM-based speech synthesis. In the task at hand we consider the adaptation task of generating a speaker model from an average voice model (AVM) and adaptation data for the desired target speaker (Yamagishi et al., 2003).

Focusing on emotional speech adaptation, it has been proved that it is very important to consider not only the means of the HMM Gaussian Distributions but also the variances. This means that it is necessary for the adaptation algorithms to be more complex, or “constrained” as it is called. Ultimately, constrained structural maximum a posteriori linear regression (CSMAPLR) has been proposed and has been proven to be extremely successful for speaker adaptation, particularly when adapting from average voice models (Yamagishi et al., 2009).

3.1.1. CSMAPLR adaptation

CSMAPLR consists in applying the structural MAP criterion (SMAP) (Shinoda and Lee, 1997) to the CMLLR adaptation algorithm (Gales, 1998) and using the recursive MAP criterion (Chesta et al., 1999) to estimate the transforms for simultaneously transforming the mean vectors and covariance matrices of the state output and duration distributions of the speaker model.

There are three main reasons for using CSMAPLR as the adaptation technique. First of all is the aforementioned capability of not only adapting the mean vectors but also the covariance matrices. The second reason that differentiates CSMAPLR from the more traditional CMLLR adaptation, is that CSMAPLR makes use of the linguistic information of the regression tree by doing recursive MAP-based estimation of the transformation matrices from the root of the context decision tree to the lower nodes, combining the advantages of SMAP and CMLLR. Finally, the fact that CSMAPLR relies on MAP-based estimations means that it is robust when using sparse adaptation data, which is frequently the case in the emotional speech synthesis task.

3.1.2. Emotion transplantation based on cascade adaptation

The concept of cascade transforms has been used previously in automatic speech recognition to adapt the background models both to the target speaker and noise at the same time (Seltzer et al., 2012). The transplantation method we present is based on the same concept, but in this case we propose chaining transformations that model both the emotional source speaker and the target speaker to produce emotional speech synthesis models, as introduced in the preliminary version of the system (Lorenzo-Trueba et al., 2013a).

In Fig. 1 we can see the block diagram representation of the proposed emotion transplantation method. If we define the CSMPALR adaptation functions in terms of their rotation matrix ζ and bias vector ϵ:

$$\overline{\mu}_{emo} = \zeta_{emo}\mu_N + \epsilon_{emo} \tag{1}$$

$$\overline{\Sigma}_{emo} = \zeta_{emo}\Sigma_N\zeta_{emo}^T \tag{2}$$

$$\overline{\mu}_{spk} = \zeta_{spk}\mu_N + \epsilon_{spk} \tag{3}$$

$$\overline{\Sigma}_{spk} = \zeta_{spk}\Sigma_N\zeta_{spk}^T \tag{4}$$

where $\overline{\mu}_{emo/spk}$ and $\overline{\Sigma}_{emo/spk}$ are the mean vectors and covariance matrices of the emotional source models and target speaker model respectively. Then, the model resulting from applying in cascade the emotion transforms and then the speaker transforms becomes:

$$\overline{\mu}_{tar} = \zeta_{spk}\zeta_{emo}\mu_N + \zeta_{spk}\epsilon_{emo} + \epsilon_{spk} \tag{5}$$

$$\overline{\Sigma}_{tar} = \zeta_{spk}\zeta_{emo}\Sigma_N\zeta_{emo}^T\zeta_{spk}^T \tag{6}$$

The resulting emotional target model will be able to produce emotional synthetic speech for the target speaker even if emotional training data for the target speaker is not available.

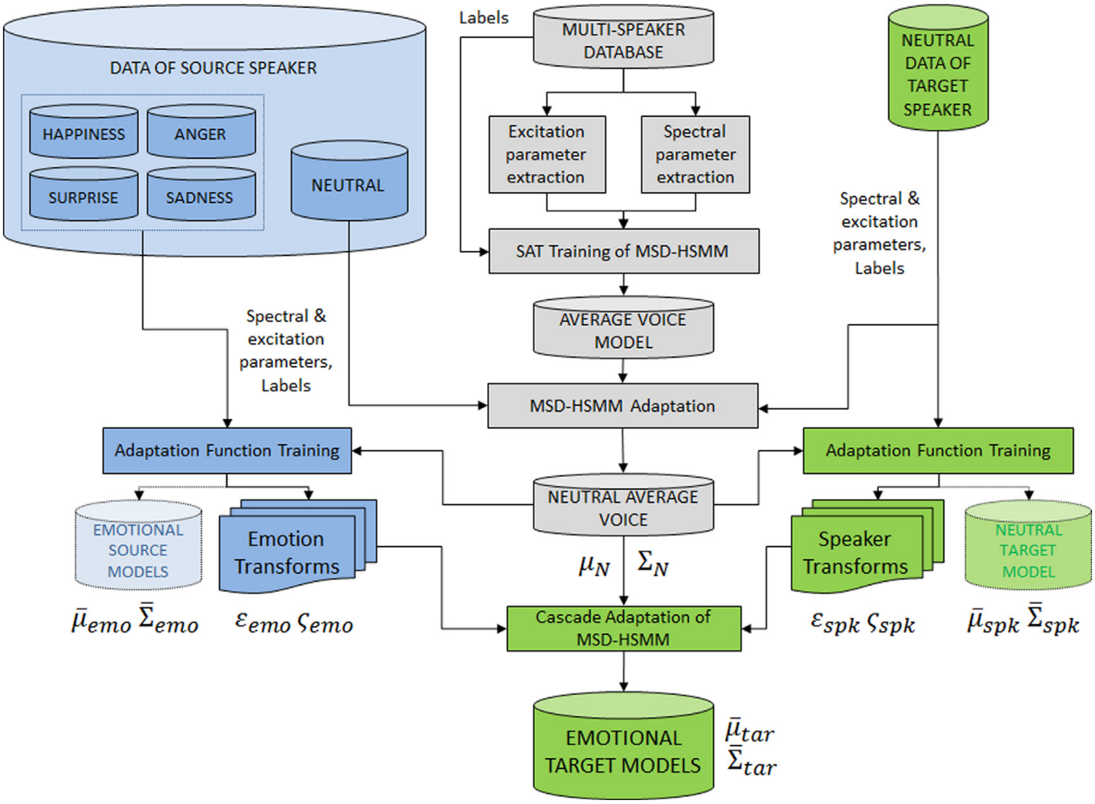
DATA OF SOURCE SPEAKER
HAPPINESS
ANGER
SURPRISE
SADNESS
NEUTRAL
Labels
MULTI-SPEAKER DATABASE
Excitation parameter extraction
Spectral parameter extraction
SAT Training of MSD-HSMM
NEUTRAL DATA OF TARGET SPEAKER
Spectral & excitation parameters, Labels
Spectral & excitation parameters, Labels
AVERAGE VOICE MODEL
MSD-HSMM Adaptation
Adaptation Function Training
Adaptation Function Training
NEUTRAL AVERAGE VOICE
EMOTIONAL SOURCE MODELS
Emotion Transforms
Speaker Transforms
NEUTRAL TARGET MODEL
$\bar{\mu}_{emo}$ $\bar{\Sigma}_{emo}$
ε_{emo} ς_{emo}
μ_N Σ_N
ε_{spk} ς_{spk}
$\bar{\mu}_{spk}$ $\bar{\Sigma}_{spk}$
Cascade Adaptation of MSD-HSMM
EMOTIONAL TARGET MODELS
$\bar{\mu}_{tar}$
$\bar{\Sigma}_{tar}$

Fig. 1. Schematic of the emotion transplantation method.

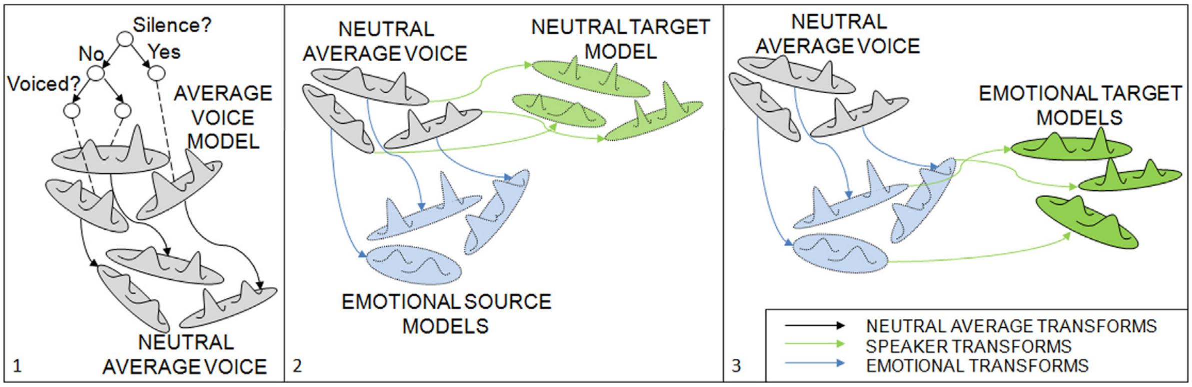

Fig. 2. Step by step block diagram of the emotion transplantation method. The ellipses represent the speaker models and the arrows the adaptation transforms.

3.2. *Proposed emotion transplantation method*

The proposed transplantation method can be summed up in three steps:

1. Adapt a neutral average voice from the average voice model to work as a reference emotion (Fig. 2(1)).
2. Adapt the neutral target model and the emotional source models from the neutral average voice (Fig. 2(2)).
3. Apply in cascade the emotion and speaker transforms to the neutral average voice. The results are the emotional target models (Fig. 2(3)).

The average voice model is obtained by applying speaker adaptive training (SAT) (Anastasakos et al., 1997) with as much training data as possible, which allows us to obtain a very context-rich background model to work with. A robust and complete AVM will be capable of producing better speech quality at synthesis time even with sparse emotional adaptation data. Also, sharing a background model for all the adaptation functions makes the cascade adaptation easier, because the context decision trees will be shared, making the adaptation functions immediately compatible between adapted models.

Adapting the neutral average voice (NAV) from the AVM is necessary because in the second step we have two objectives: on one hand we want to be able to learn the differences between the emotional source model and the NAV, effectively learning the nuances of the desired emotional speech. On the other hand we want to learn the difference between the target speaker speaking in a typical, neutral style and the NAV, thus learning the nuances of the target speaker identity. If both adaptation functions are not obtained from a common reference emotion, neutral speech in the present case, they will not learn the desired characteristics and the transplantation process would not be successful. Ideally, we want to have both data for the target emotion and reference emotion for the same speaker so the emotion adaptation function defines purely the emotional source model, but if that is not available we can assume that using an average of different speakers will show the relevant information of the emotions while lessening the identities of the speakers.

Finally, we apply in cascade the emotion and speaker transforms as defined previously (Fig. 2(3)), obtaining the desired emotional target models. The produced emotional strength can be easily controlled by means of a transplantation control ratio to linearly scale the adaptation function. Neither the target speaker nor the target emotion data had to be present in the AVM, and in the presented evaluation (Section 4) we prove it to be successful with as little as 5 min of target speaker speech data or 30 min of emotional speech, which is why the proposed transplantation method is a good way of providing scalability in expressive speech synthesis. Nonetheless, in order to provide a basis for comparison, we also propose an emotion transplantation baseline.

3.2.1. Transplantation baseline: transplanting into an average emotion

The proposed emotion transplantation baseline joins all our emotional data into an average emotion model (Qin et al., 2006) to be used as the emotional source. This average, when transplanted into the neutral target model, can be expected to imbue an undefined emotion that removes the typical monotony in read speech models. This alternative can also be expected to provide higher quality speech when compared to transplanting a single emotion as the adaptation process for the average emotion can make use of much more data, thus giving more stable adaptation functions. This approach could be very useful when the task does not require us to synthesize any particular emotion or if we do not have enough emotional data to obtain good emotion transplantation quality.

Naturally, this alternative also presents numerous shortcomings: if there is a significant bias toward positive or negative data in the average emotion model, transplanting the average emotion could be the same as transplanting an emotion, resulting on unnatural synthesized utterances for opposite emotions such as producing happy speech for a sad text. Another expected problem is that the naturalness should be lower than transplanting the correct emotion for the text to be synthesized.

4. Perceptual evaluation

The goal of the perceptual evaluation was to verify if the expressiveness was transplanted successfully in terms of naturalness, speech quality and emotional strength. Naturalness measure was done by means of forced preference tests, as they are very useful when we want to compare systems that are similar between them but with variation in some conditions (King et al., 2008). Three different evaluations were carried out: a first evaluation that compared the average emotional model transplantation against the traditional neutral synthetic voices in order to establish an emotion transplantation baseline. A second evaluation to compare the proposed emotion transplantation system against neutral voices so as to validate the usefulness of the presented methodology. A third evaluation comparing the average emotion transplantation with the proposed emotion transplantation system, aiming to rate the performance of the proposed system against the baseline.

Table 2
Results for the preliminary evaluation on transplanted emotion identification (Lorenzo-Trueba et al., 2013a), emotional strength and similarity to the target speaker for a transplantation control ratio of 1.00. Results are relative to those of the natural voice.

Metric	System	Anger (%)	Happiness (%)	Sadness (%)	Surprise (%)
Emotion id.	Proposed	36.0	52.1	88.5	73.1
Emotional strength	Proposed	65.2	69.1	67.7	66.4
Similarity	Proposed	84.6	94.5	86.2	67.8

Table 3
Sample emotional utterances used for the proposed preference evaluations translated from Spanish.

Emotion	Utterance
Anger	I have just picked up my clothes from the laundry, they have been ruined!
Happiness	I am so happy! I passed the exam in the first try even though the subject was so hard.
Sadness	It has been two years since he passed away, we miss him dearly.
Surprise	You cannot believe who I just saw walking down the street!

4.1. Evaluations design

Four emotions (anger, happiness, sadness and surprise) learned from the SEV corpus were transplanted into 3 male speakers and 3 female speakers, so for each test the total number of systems was 24. Following the Latin-square (Gao, 2005) approach this meant that we needed 24 different utterances to be synthesized for all the systems to be presented to the listeners in a random order without repetitions. In the end we decided for considering 24 utterances per evaluation and not multiples of it so as to keep the evaluation itself from becoming too long, lasting around 30 min in this implementation. In the test, two audio samples were presented to the listener by means of a web interface: the transcription of the synthesized texts and the intended emotion to be transmitted. The samples could be played as many times as desired by the listener. The synthesized texts, not present in the training data, were carefully designed by ourselves to present clear emotional context. A sample utterance for each emotion can be seen in Table 3. The listener was asked their preference for which of the samples was more adequate to transmit the desired emotion, being forced to pick one or the other. Then they were asked to rate the utterances in the traditional 5 point MOS evaluation for both speech quality (from very bad to very good) and emotional strength (very low to very high). The evaluation for speech quality and emotional strength had to be answered for both samples regardless of the selected preference. In the case of the multiple choice secondary evaluation (Section 4.3.1), the listeners were presented the transcribed synthesized text, the target emotion and a set of 5 utterances and were asked to select one sample out of the five that they believed better conveyed the target emotion. The 5 utterances corresponded to the 5 emotions (anger, happiness, neutral, sadness, and surprise) and were randomized in order for each sample.

The carried out evaluations did not include an analysis of similarity and emotional strength control despite them being advantages of the proposed systems because they have been thoroughly analyzed in previous tests. Our initial emotion transplantation evaluation (Lorenzo-Trueba et al., 2013a) measured the similarity, emotion recognition rates, speech quality and emotional strength of the proposed systems with different emotion transplantation control ratios, using neutral synthetic speech as baseline and natural voice as a top-line. Only the male source and target speakers of the introduced corpora (Section 2) were used for this preliminary evaluation. The results for the emotion identification rates can be seen in Table 2. It must be noted that because the synthesized text in the preliminary evaluation was designed to be completely neutral, the emotion identification task was significantly harder than for the task in the present evaluation, where not only the text was carefully designed by us to be clearly emotional but also we provided the evaluators with the intended emotion. The previous test showed how that the attainable similarity is comparable to natural speech, and that values of the transplantation control ratio that keep a good balance between speech quality and emotional strength range between 0.75 and 1.00. This was because higher ratios degraded the speech quality too much and lower ratios limited too much the emotion identification rates and the perceived emotional strength. Consequently, for the present evaluation we did not measure similarity any further, and the applied emotion transplantation control ratio was 1.00. Regarding the statistical significance of the results, for the preference test we applied the Chi-squared

Table 4
Source and target data used for training the adaptation functions for the 3 systems evaluated.

System	Neutral source	Emotional source	Neutral target
Neutral	JOA/ROS	None	All 6 target speakers
Average	JOA/ROS	All 4 emotions	All 6 target speakers
Proposed	JOA/ROS	Target emotion	All 6 target speakers

Table 5
Results for the first evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant.

Metric	System	Anger	Happiness	Sadness	Surprise
Preference	Average	66%	87%	61%	84%
	Neutral	34%	13%	39%	16%
Speech quality	Average	3.11	3.19	3.16	3.21 (X)
	Neutral	3.38	3.38	3.38	3.38 (X)
Emotional strength	Average	3.37	3.33	2.93	3.34
	Neutral	2.57	2.57	2.57	2.57

criterion and for the speech quality and emotional strength MOS tests we applied the Wilkoxon Signed-Rank Test for a 95% confidence ratio. A minimum of 24 subjects per evaluation was fixed in order to guarantee the full coverage of the Latin square matrix, totaling at least 24 samples per speaker, emotion and evaluation.

Regarding the speaker models, the three considered systems (neutral, average and proposed) both share a common AVM trained with three feature streams with their Δ and Δ^2 coefficients: logarithm of the fundamental frequency (1 coefficient), mel-cepstral analysis coefficients (MCEP, 60 coefficients) and aperiodicity bands (25 coefficients). For the synthesis part we used the STRAIGHT vocoder (Kawahara et al., 2001) for all the systems. The data used to train each system can be seen in Table 4.

4.2. Evaluation 1: average emotion transplantation system against neutral system

The first test aims to establish the evaluation baseline by comparing the average emotion transplantation system and the traditional neutral synthesis system, and was carried out by 28 listeners. The first aspect to note is that the preference test greatly favors the average system as we can see in Table 5, with an average preference rate of 75%. The results for emotional strength also show that the average system is perceived generally as more expressive, with an average of 0.7 points more perceived strength than in the neutral system, especially in the angry and surprised cases. Finally, speech quality suffers a slight degradation of 0.2 points in average, being this category the only one with non significant differences (Fig. 3). All the boxplots shown in this paper follow the same structure: the boxes capture the ranges between the first and third quartile, with the median represented by the thick line and the whiskers showing the highest and lowest datum within 1.5 the inter-quartile range.

4.3. Evaluation 2: proposed emotion transplantation system against neutral system

The second test looks to show the advantages of the proposed emotional target model when compared against the neutral target model, and it was carried out by 27 listeners. The results (Table 6) show a clear preference for the transplanted system, in this case in an average of 88% preference rate. This rate is considerably higher than in the first evaluation, and reaches preference rates as high as the 96% or 95% for the happy and surprised systems respectively. This result is further proven by the emotional strength results, where there is an average of 1.2 points in the perceived emotional strength, especially once again in the happy and surprised systems, clearly shown in the boxplots (Fig. 4). Speech quality, on the other hand, suffers a slightly stronger degradation (an average of 0.4 points in this case).

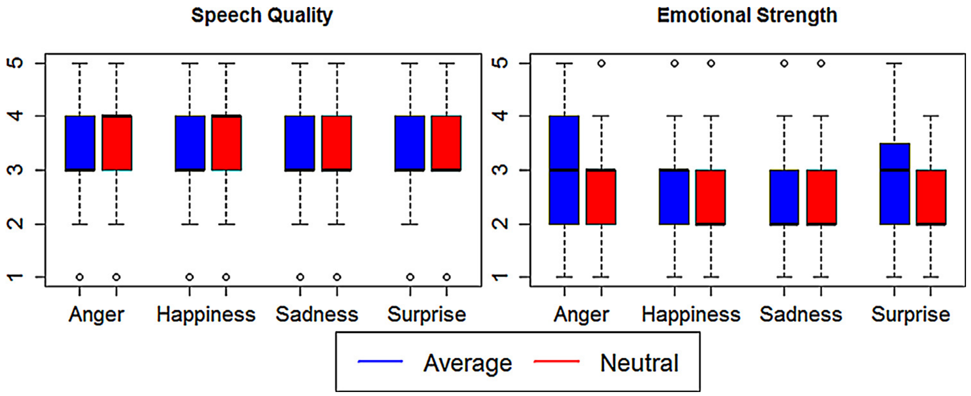

Fig. 3. Boxplots for the results of both speech quality and emotional strength for the first evaluation. The blue bars represent the average emotion transplantation system and the red bars represent the neutral system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Results for the second evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant.

Metric	System	Anger	Happiness	Sadness	Surprise
Preference	Proposed	77%	96%	82%	95%
	Neutral	23%	4%	18%	5%
Speech quality	Proposed	3.05 (X)	2.81	2.70	3.05
	Neutral	3.30 (X)	3.30	3.30	3.30
Emotional strength	Proposed	3.03	3.61	3.26	3.73
	Neutral	2.17	2.17	2.17	2.17

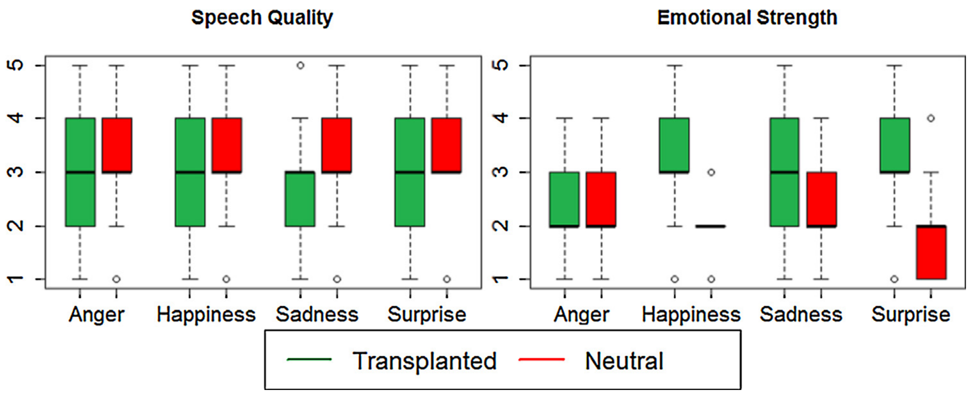

Fig. 4. Boxplots for the results of both speech quality and emotional strength for the second evaluation. The green bars represent the proposed transplantation system and the red bars represent the neutral system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3.1. Evaluation 2.2: proposed emotion transplantation system against neutral system multiple choice preference

In order to measure the capabilities of the proposed emotion transplantation system to convey the desired target emotion we carried out a follow up to the second test, where we asked the evaluators to pick the utterance they felt that conveyed more clearly the target emotion out of a pool of 5 (anger, happiness, neutral, sadness and surprise). This test was carried out by 25 listeners and the confusion matrix resulting can be seen in Table 7.

The results show how the evaluators significantly prefer the utterances that were transplanted with the desired emotion when compared to the rest of the utterances with an average preference of around 70% for the intended one, which proves that the method is capable of adequately transplanting emotions. In particular we can see how there

Table 7
Confusion matrix for the multiple choice preference test considering all four transplanted emotions and neutral voice. Rows show the chosen preferred system and columns the actual transplanted emotion.

Preference	Anger (%)	Happiness (%)	Sadness (%)	Surprise (%)
Anger	67.9	2.6	14.2	3.4
Happiness	14.7	71.1	3.2	20.1
Sadness	1.3	2.0	70.3	3.4
Surprise	14.7	23.7	4.5	71.8
Neutral	1.3	0.7	7.7	1.3

Table 8
Results for the third evaluation averaged across speakers for the three categories. The X represents results that are not statistically significant.

Metric	System	Anger	Happiness	Sadness	Surprise
Preference	Proposed	57%	80%	76%	84%
	Average	43%	20%	24%	16%
Speech quality	Proposed	3.22 (X)	3.06	3.07 (X)	3.12
	Average	3.13 (X)	2.94	3.06 (X)	2.97
Emotional strength	Proposed	3.10	3.24	3.12	3.35
	Average	2.88	2.67	2.50	2.66

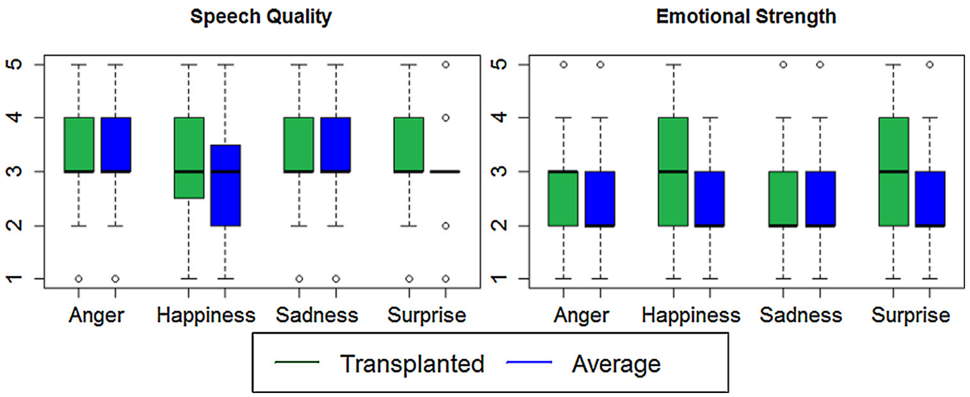

Fig. 5. Boxplots for the results of both speech quality and emotional strength for the third evaluation. The green bars represent the proposed transplantation system and the blue bars represent the average emotion transplantation system.

was almost no preference for the neutral utterances, even in the sad system where a neutral way of speaking could sometimes be acceptable. Also, the confusion between happiness and surprise makes sense, with both of them being positive emotions with high arousal that tend to produce higher pitched voices. The confusion of anger with happiness and surprise can be explained by the fact that sometimes anger is conveyed by a high arousal variation called hot anger. Sadness on the other hand is slightly confused with both cold anger and neutral, both low arousal states with slower speaking rhythm and lower pitch.

4.4. *Evaluation 3: proposed emotion transplantation system against average emotion transplantation system*

The third and final evaluation aims to measure the advantages of the proposed transplantation system against the baseline average emotion transplantation system, which means measuring the advantages of transplanting the emotion associated to the text to be synthesized against using an average standard emotion for all texts. In this case, the results obtained from 31 listeners, clearly show a preference for the proposed transplantation system, with an average preference rate of 75% (Table 8). Also, the perceived emotional strength results show an increase of 0.5 points when selecting the adequate emotion, stronger for the happy and surprised systems as shown in the boxplot (Fig. 5), but also

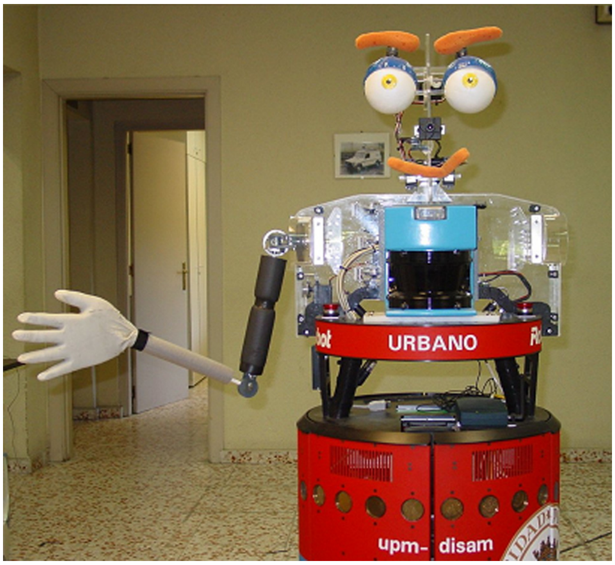

Fig. 6. Picture of the robotic assistant, Urbano.

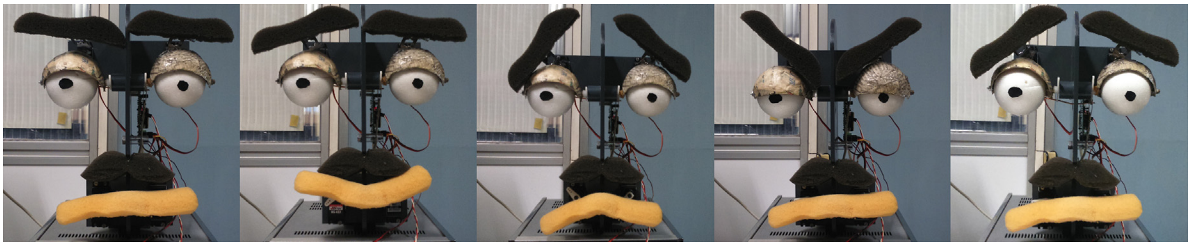

Fig. 7. Expressive faces for Groucho in the following order: neutral, happiness, sadness, anger and surprise.

present in the angry and sad systems. Finally, for speech quality we see extremely close results, less than 0.1 points in average.

5. Evaluation of the transplantation system in a real robotic application

In order to evaluate the improvements attainable by the proposed emotion transplantation method, we also designed an evaluation in a more realistic application: synthesizing the voice of a robotic laboratory assistant (Fig. 6). This robotic laboratory assistant, in its finished form, would be capable of moving around a laboratory interacting with the students and answering their questions or doing quizzes to evaluate their progress. For this particular evaluation, as we wanted to focus on how we could improve the robot's interaction capabilities by adapting its way of speaking to the context, we only used the robot's head in a 'Wizard of Oz' dialog scenario. Our role as the 'Wizard of Oz' was to replace the speech recognition system, starting the robot's turns in the scripted dialog.

5.1. Robot description

We have developed a face with 5 degrees of freedom, 2 to control the mouth, one for each eyebrow and another one for closing both eyes, similar to (Rodriguez-Losada et al., 2008). The eyes were completely static, the mouth could not be opened and both eyelids had to be moved together. The actuators are based on Futaba servomotors and a simple servo-controller board allows sending sequences of control bytes through a serial port from a portable PC on-board the robot. As shown in Fig. 7, it is perfectly capable of showing basic emotions such as happiness, sadness, anger or surprise.

Table 9
Sample happiness dialog used in the robot evaluation, translated from Spanish. The italics text shows the emotional turns in the dialog.

Turn	Robotic assistant	Student
1	Good morning, can I help you?	Yes, I wanted to measure this, but something is wrong.
2	Have you tried measuring with the second channel?	Oh! That was the problem! Thank you very much.
3	*Great! I was glad to be of help.*	

Table 10
Results for the first preference evaluation with the robotic assistant. The X represents results that are not statistically significant.

Metric	System	Anger (%)	Happiness (%)	Sadness (%)	Surprise (%)
Preference	Proposed	77	83	54 (X)	88
	Neutral	23	17	46 (X)	13

5.2. Robot evaluation design

So as to emulate a realistic scenario, we developed a total of 36 interaction scenarios (12 for happiness, 12 for sadness, 6 for anger and 6 for surprise) of which 6 would be carried on by each evaluator (2 for happiness, 2 for sadness, 1 for anger and 1 for surprise). A sample dialog translated to English can be seen in Table 9. This dialog distribution was chosen because we believe that it is more common in a student-robotic assistant dialog to show happiness and sadness than anger or surprise. Regardless, all of them had to be present to fully test the proposed system. This evaluation was also designed with the Latin square principle in mind, so that the order of the emotions and dialogs were fully randomized. In each scenario a short dialog between 1 and 3 interaction turns would be played out between the robot and the evaluator, and it would be played twice (once for the neutral system and once for the proposed emotional transplantation system, in a random order). The robotic face would then adjust its expression to match that of the context regardless of the system that was being used, so that the only variation between the iteration of the dialogs was the synthetic speech. Also, to prevent the evaluation from becoming too long, we only considered one of our target speakers, UVD.

The evaluators, which were real students from UPM University, were explained that they were going to evaluate a robotic laboratory assistant called Groucho that wanted to be capable of adapting its way of speaking according to the context. Then they were given the scripts for the dialogs and were told that each dialog would be played twice, once for each emotional system. After carrying out each scenario twice, the evaluator was asked to choose between both versions which one they felt was better according to the dialog that was played out in a forced preference test. If they were undecided, the scenarios could be carried out multiple times, but only the complete scenario.

5.3. Robot evaluation results

The results for the 24 students that took part on a first evaluation can be seen in Table 10. There we see how the preference is heavily in favor of the proposed emotion transplantation system, which clearly shows that being able to provide a speech synthesis system capable of adapting its speaking style (emotions in this particular application) to the context can provide significant increases in the satisfaction of the users of the dialog system. It is only sadness that did not provide significant improvements, and informal dialog with the evaluators that disliked it showed that the problem was that they felt that the sad system was out of place in the laboratory assistant. Careful inspection of the source sadness data showed that it presents a tearful tone, which adds an aspect to the sad emotion that makes it inadequate for this particular domain.

In an attempt to fix the problems with sadness we prepared a second evaluation in which sadness was transplanted with only a 0.5 transplantation control ratio so that the synthesized emotional strength was lessened, while the rest of the emotions remained at 1.0. In this second iteration of the system 12 students participated, giving results very similar to the previous iteration as we can see in Table 11. Informal conversation once again showed that while less recognizable, the effects of tearfulness still made the transplanted sadness to feel out of place, showing that reducing

Table 11
Results for the second preference evaluation with the robotic assistant. The X represents results that are not statistically significant.

Metric	System	Anger (%)	Happiness (%)	Sadness (%)	Surprise (%)
Preference	Proposed	83	83	46 (X)	83
	Neutral	17	17	54 (X)	17

the emotional strength is not enough to attenuated the tearfulness aspect from the sad models as they are conveyed by complex spectral cues that cannot be eliminated without removing also sadness itself.

6. Discussion

To sum up, the baseline average emotion transplantation system provides an average preference rate with an increase of 0.7 points in emotional strength at the cost of an average of 0.2 points in speech quality. In comparison, the proposed emotion transplantation system provides an average 87% preference rate increasing the perceived emotional strength in an average of 1.2 points at the cost of 0.4 points in speech quality. The final test, comparing the baseline and proposed systems reinforces the idea that transplanting the correct emotion is a better approach, as there is a preference for the proposed system of 74%, with a perceived increase in 0.5 points in emotional strength, with an only partially statistically significant decrease in speech quality smaller than 0.1 points. All in all, we can say that the proposed emotion transplantation method is clearly capable of imbuing the emotional information learned from a source speaker into different target speakers regardless of gender with significant increases in perceived naturalness and emotional strength when compared to traditional systems at a slight cost in speech quality. In total we had 86 evaluators, 28 for the first, 27 for the second and 31 for the third and last one. This meant a grand total of 2046 evaluated utterances and 516 per emotion, providing statistically significant results in the presented evaluations.

In order to quantify the robustness of the proposed transplantation system to different speaker conditions we have carried out a number of ANOVA and correlation analyses of the effect that speaker gender, speaker identity and neutral synthetic model speech quality have on the speech quality and preference of the proposed system. The results of the ANOVA analysis show that neither speaker gender nor speaker identity have any relevant effect on the transplanted speech quality or preference. On the other hand, the speech quality of the neutral model clearly impacts the transplanted speech quality and preference. This impact is further confirmed by the Pearson's product-moment correlation between neutral and transplanted speech quality, taking a value of 0.38. Based on these results we can conclude that the proposed transplantation system is significantly robust against speaker variability, with the attainable speech quality strongly depending on the quality of the quality of the source model. This means we can provide high quality emotional speech synthesis for any speaker as long as we have a high quality neutral speech of the target speaker.

Finally, we carried out a pair of evaluations which considered the interaction between an end user and a robot. These evaluations show that the students clearly prefer when the robot is capable of adapting its way of speaking to the situation, and an informal questionnaire after the evaluations proved that the overall satisfaction with the emotional system was much higher. In the end this shows that there is a lot to gain from implementing expressive and adaptive versions of human–machine interaction systems. This evaluation has also proven how when dealing with specific application domains one has to consider not only the quality of the emotion but also the appropriateness. In our particular case the listeners felt that sadness did not fit the laboratory assistant task, so it would have been better to replace it with a neutral voice.

7. Conclusions

We have proposed an emotion transplantation method capable of learning the paralinguistic nuances of any particular emotion in order to transplant them into a new target speaker for whom only neutral read speech recordings are available. This is done by means of chaining a pair of CSMAPLR adaptation functions, one that characterizes the target speaker identity and another that captures the paralinguistic characteristics of the desired emotion. Finally a triplet of perceptual evaluations were carried out.

For the perceptual evaluations, four emotions (anger, happiness, sadness and surprise) from a Spanish emotional database and six target speakers (three male and three female) were considered. A first evaluation compared in terms of naturalness, speech quality and emotional strength a baseline average emotion transplantation with traditional neutral read speech synthesis. The second evaluation compared the proposed emotion transplantation method with neutral speech synthesis, and the final evaluation compared the baseline with the proposed method. The results clearly proved how emotion transplantation greatly increases the perceived emotional strength of the synthesized utterances (up to 1.2 points for the proposed method), also showing a very significant increase in preference (an average of 87% for the proposed system) at a cost of only 0.4 points in speech quality. This results are enforced by the average 70% preference rates in the multiple-choice evaluation. The comparisons also showed how transplanting the correct emotion is a much better alternative than transplanting an average emotion, with significant increases of 0.5 points in perceived emotional strength and a 74% preference rate at a cost of less than 0.1 points in speech quality. We have also seen that there is no correlation between speaker gender or identity and the obtained results, while there is a strong correlation between the source neutral system speech quality and the transplanted quality, meaning that the proposed transplantation system is robust against speaker variability, and is capable of providing high quality emotional models if the source neutral speaker model is of high quality. The transplanted voices provided us with multiple emotional voices for the robots developed in the ARABOT and INAPRA projects, but by imbuing emotions on previously available speakers recorded in just a neutral speaking style.

Future work includes increasing the speech quality obtained through the transplantation process, verifying the method with different expressiveness such as speaking styles, fine-tuning and evaluating the transplanted emotions (and the whole emotional robotic system) in real scenarios such as the Principe Felipe Science Museum, and considering the possibility of applying it across languages to transplant effects known to be language independent such as Lombard speech.

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 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. The authors also want to thank the Multimedia Technology Group from Universidad de Vigo and the Speech Processing Group (VEU) from Universidad Politécnica de Cataluña for sharing some of the databases used in this research.

References

Adell, J., Bonafonte, A., Escudero-Mancebo, D., 2010. Modelling filled pauses prosody to synthesise disfluent speech. In: Proc. ISCA Speech Prosody, Chicago, USA.

Adell, J., Escudero, D., Bonafonte, A., 2012. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. Speech Commun. 54, 459–476.

Anastasakos, T., McDonough, J., Makhoul, J., 1997. Speaker adaptive training: a maximum likelihood approach to speaker normalization. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, ICASSP-97, IEEE vol. 2, pp. 1043–1046.

Andersson, S., Georgila, K., Traum, D., Aylett, M., Clark, R.A., 2010. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In: Speech Prosody.

Andersson, S., Yamagishi, J., Clark, R.A., 2012. Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. Speech Commun. 54, 175–188.

Barra-Chicote, R., Ph.D. thesis 2011. Contributions to the Analysis, Design and Evaluation of Strategies for Corpus-based Emotional Speech Synthesis. ETSIT-UPM.

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., 2008. Spanish expressive voices: corpus for emotion research in Spanish. In: Proc. of LREC.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M., Macias-Guarasa, J., 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Commun. 52, 394–404.

Bonafonte, A., Moreno, A., 2008. Documentation of the UPC-ESMA Spanish database. In: TALP Research Center. Universitat Politecnica de Catalunya, Barcelona, pp. 2781–2784.

Chen, L., Gales, M., Wan, V., Latorre, J., Akamine, M., 2012. Exploring rich expressive information from audiobook data using cluster adaptive training. In: Interspeech 2012, 13th Annual Conference of the International Speech Communication Association, September 9–13, Portland, Oregon.

Chesta, C., Siohan, O., Lee, C.-H., 1999. Maximum a posteriori linear regression for hidden Markov model adaptation. In: Eurospeech.

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn. 44, 572–587.

Erro, D., Navas, E., Herndez, I., Saratxaga, I., 2010. Emotion conversion based on prosodic unit selection. IEEE Trans. Audio Speech Lang. Process. 18, 974–983.

Banga C.M., E.T., 2010. Documentation of the UVIGO-ESDA Spanish database. In: Technical Report Grupo de Tecnoloxias Multimedia. Universidad de Vigo, Vigo, España.

Gales, M.J., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang., 12.

Gales, M.J., 2000. Cluster adaptive training of hidden Markov models. IEEE Trans. Speech Audio Process. 8, 417–428.

Gao, L., 2005. Latin Squares in Experimental Design. Michigan State University.

Hsu, C.-Y., Chen, C.-P., 2012. Speaker-dependent model interpolation for statistical emotional speech synthesis. EURASIP J. Audio Speech Music Process. 2012, 1–10.

Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. Proc. MAVEBA, 13–15.

King, S., Clark, R.A., Mayo, C., Karaiskos, V., 2008. The Blizzard Challenge., pp. 2008.

Latorre, J., Wan, V., Gales, M.J., Chen, L., Chin, K., Knill, K., Akamine, M., 2012. Speech factorization for HMM-TTS based on cluster adaptive training. In: INTERSPEECH.

Lorenzo-Trueba, J., Barra-Chicote, R., Raitio, T., Obin, N., Alku, P., Yamagishi, J., Montero, J.M., 2012, September. Towards glottal source controllability in expressive speech synthesis. In: Interspeech 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, pp. 9–13.

Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., Montero, J., 2013a. Evaluation of a transplantation algorithm for expressive speech synthesis. In: Proceedings of Workshop en Tecnologias Accesibles, IV Congreso Espa.

Lorenzo-Trueba, J., Barra-Chicote, R., Yamagishi, J., Watts, O., Montero, J.M., 2013b. Towards speaking style transplantation in speech synthesis. In: 8th ISCA Speech Synthesis Workshop.

Lutfi, S.L., Fernández-Martí nez, F., Lorenzo-Trueba, J., Barra-Chicote, R., Montero, J.M., 2013. I feel you: the design and evaluation of a domotic affect-sensitive spoken conversational agent. Sensors 13, 10519–10538.

Mendez Pazo, F., Docio Fernandez, L., Arza Rodriguez, M., Campillo Diaz, F., 2010. The albayzyn 2010 text-to-speech evaluation. In: Proceedings of VI Jornadas en Tecnologia del Habla and II Iberian SLTechWorkshop, Vigo, Spain.

Nose, T., Kobayashi, T., 2013. An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. Speech Commun. 55 (2), 347–357.

Obin, N., Lanchantin, P., Lacheret, A., Rodet, X., et al., 2011. Discrete/continuous modelling of speaking style in HMM-based speech synthesis: design and evaluation. In: Interspeech.

Picart, B., Drugman, T., Dutoit, T., 2011. Continuous control of the degree of articulation in HMM-based speech synthesis. In: INTERSPEECH, pp. 1797–1800.

Qin, L., Ling, Z.-H., Wu, Y.-J., Zhang, B.-F., Wang, R.-H., 2006. HMM-based emotional speech synthesis using average emotion model. In: Huo, Q., Ma, B., Chng, E.-S., Li, H. (Eds.), Chinese Spoken Language Processing, vol. 4274 of Lecture Notes in Computer Science. Springer Berlin, Heidelberg, pp. 233–240.

Raitio, T., Suni, A., Vainio, M., Alku, P., 2013. Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. Comput. Speech Lang.

Rodriguez-Losada, D., Matia, F., Galan, R., Hernando, M., Montero, J., Lucas, J., 2008. Urbano, an interactive mobile tour-guide robot. In: Seok, H. (Ed.), Advances in Service Robotics. In-Teh, pp. 229–252.

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010. Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. 1, 119–131.

Seltzer, M.L., Acero, A., 2012. Factored adaptation using a combination of feature-space and model-space transforms. In: INTERSPEECH.

Shinoda, K., Lee, C.-H., 1997. Structural map speaker adaptation using hierarchical priors. In: IEEE Workshop on Automatic Speech Recognition and Understanding, 1997, IEEE Proceedings, pp. 381–388.

Takeda, S., Kabuta, Y., Inoue, T., Hatoko, M., 2013. Proposal of a Japanese-speech-synthesis method with dimensional representation of emotions based on prosody as well as voice-quality conversion. Int. J. Affect. Eng. 12, 79–88.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. 17, 66–83.

Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2005. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. IEICE Trans. Inform. Syst. 88, 502–509.

Yamagishi, J., Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T., 2003. A training method of average voice model for HMM-based speech synthesis. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 86, 1956–1963.

Yanagisawa, K., Latorre, J., Wan, V., Gales, M.J., King, S., 2013. Noise robustness in HMM-TTS speaker adaptation. Order 5, 10.

Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S., 2004. Towards emotional speech synthesis: a rule based approach. Fifth ISCA Workshop on Speech Synthesis.

Jaime Lorenzo-Trueba is a Ph.D. student of the Electronic Engineering Department at Universidad Politécnica de Madrid and researcher of the Speech Technology Group. Jaime has received a M.S. degree in electronics and photonics from Keio University in Japan and a M.S. degree in Telecommunication Engineering from UPM. His main research areas are parametric speech synthesis, expressive speech synthesis and speaking styles characterization. Jaime has also been a visitor in 2014 in the National Institute of Informatics in Japan.

Roberto Barra-Chicote is an Assistant Professor at Universidad Politécnica de Madrid. He received (with highest distinction) his MSEE degree from Technical University of Madrid in 2005 and his Ph.D. degree in 2011. In 2006 he was a visitor researcher of the Center for Spoken Language Research (CSLR) at Colorado University. In 2008 he was a visitor researcher of the Centre for Speech Technology Research (CSTR) at Edinburgh University. He has research interests in speech synthesis, speech recognition and speaker diarization.

Rubén San-Segundo received the Ph.D. degree (with highest distinction) from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2002. He did two research stays in The Center of Spoken Language Research (CSLR), University of Colorado at Boulder, as visiting student. From September 2001 through February 2003, he worked at the Speech Technology Group of the Telefónica I+D. Currently, he is an Associate Professor in the Department of Electronic Engineering at UPM. He has been the Coordinator of the Spanish Network on Speech Technologies and the vice-chair of the Special Interest Group of ISCA on Iberian Languages.

Javier Ferreiros received the M.S.E.E. and Ph.D. degrees with highest distinctions from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1990 and 1996, respectively. Since 1988, he has been a member of the Speech Technology Group at UPM, where he holds an Associate Professor position. He has been the Director for academic planning of the Escuela Técnica Superior de Ingenieros de Telecomunicación. From October 1999 to April 2000, he stayed at ICSI, Berkeley, CA, as a Visiting Researcher. His research interests focus on spoken dialog systems.

Junichi Yamagishi is a senior research fellow in CSTR at the University of Edinburgh. He is also an associate professor of National Institute of Informatics in Japan. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for pioneering speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis. He was awarded the Itakura Prize for his achievements in adaptive speech synthesis and the 2012 Kiyasu Special Industrial Achievement Award. In 2012 he was area chair for Interspeech and elected to membership of the IEEE Signal Processing Society Speech & Language Technical Committee.

Juan M. Montero is an Associate Professor of the Electronic Engineering Department at Universidad Politécnica de Madrid and researcher of the Speech Technology Group. Dr. Montero received a M.S. degree in Telecommunication Engineering and a Ph.D. degree "cum laude" from UPM, and was awarded the La Caixa COIT Prize as the best Ph.D. thesis. His main research areas are parametric speech synthesis, affective computing, speaking style modeling and Project Based Learning. Dr. Montero has been visiting researcher at the ICSI (Berkeley), DFKI (Saarbrucken) and CSTR (Edinburgh).