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
0885-2308/

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).