Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 278–294

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Shape-based modeling of the fundamental frequency contour for emotion detection in speech

Juan Pablo Arias [a], Carlos Busso [b], Nestor Becerra Yoma [a,*]

[a] *Speech Processing and Transmission Laboratory, Department of Electrical Engineering, Universidad de Chile, Santiago, Chile*
[b] *Multimodal Signal Processing Laboratory, The University of Texas at Dallas, Richardson, TX 75080, USA*

Received 20 July 2012; received in revised form 25 March 2013; accepted 16 July 2013
Available online 26 July 2013

Abstract

This paper proposes the use of neutral reference models to detect local emotional prominence in the fundamental frequency. A novel approach based on *functional data analysis* (FDA) is presented, which aims to capture the intrinsic variability of F0 contours. The neutral models are represented by a basis of functions and the testing F0 contour is characterized by the projections onto that basis. For a given F0 contour, we estimate the functional *principal component analysis* (PCA) projections, which are used as features for emotion detection. The approach is evaluated with lexicon-dependent (i.e., one functional PCA basis per sentence) and lexicon-independent (i.e., a single functional PCA basis across sentences) models. The experimental results show that the proposed system can lead to accuracies as high as 75.8% in binary emotion classification, which is 6.2% higher than the accuracy achieved by a benchmark system trained with global F0 statistics. The approach can be implemented at sub-sentence level (e.g., 0.5 s segments), facilitating the detection of localized emotional information conveyed within the sentence. The approach is validated with the SEMAINE database, which is a spontaneous corpus. The results indicate that the proposed scheme can be effectively employed in real applications to detect emotional speech.

Keywords: Emotion detection; F0 contour modeling; Emotional speech analysis; Expressive speech

1. Introduction

Emotional understanding is a crucial skill in human communication. It plays an important role not only in interpersonal interactions, but also in many cognitive activities such as rational decision making, perception and learning (Picard, 1997). For this reason, modeling and recognizing emotions is essential in the design and implementation of *human-machine interfaces* (HMIs) that are more in tune with the user's needs. Systems that are aware of the user's emotional state will facilitate several new scientific avenues that serve as truly innovative advancements in security and defense (e.g. threat detection), health informatics (e.g., depression, autism), and education (e.g., tutoring system) (Burleson and Picard, 2004; Langenecker et al., 2005). Given the important role of speech in the expression of emotions, an increasing number of publications have reported progress in automatic emotion recognition and detection using acoustic features. Complete reviews are given by Cowie et al. (2001), Zeng et al. (2009), Schuller et al. (2011a), Koolagudi and Rao (2012), El Ayadi et al. (2011).

[*] Corresponding author. Tel.: +56 2 29784205; fax: +56 2 26953881.
E-mail addresses: nbecerra@ing.uchile.cl, nbecerray@gmail.com (N.B. Yoma).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.07.002

The dominant approach in emotion recognition from speech consists in estimating global statistics or functionals at sentence level from low level descriptors such as F0, energy and *Mel-frequency cepstral coefficients* (MFCCs) (Schuller et al., 2011a). Among prosodic based features, gross pitch statistics such as mean, maximum, minimum and range are considered as the most emotionally prominent parameters (Busso et al., 2009). One limitation of global statistics is the assumption that every frame in the sentence is equally important. Studies have shown that emotional information is not uniformly distributed in time (Lee et al., 2004; Busso and Narayanan, 2007). For example, the intonation in happy speech tends to increase at the end of the sentence (Wang et al., 2005). Since the statistics are computed at the global level, it is not possible to identify local salient segments or focal points within the sentence. Furthermore, features describing global statistics do not capture local variations (e.g., in F0 contours), which in turn could provide useful information for emotion detection. In this context, this paper proposes a novel shape-based approach to detect emotionally salient temporal segments in the speech using *functional data analysis* (FDA). The detection of localized emotional segments can shift current approaches in affective computing. Instead of recognizing the emotional content on pre-segmented sentences, the problem can be formulated as a detection paradigm, which is appealing from an application perspective (e.g., continuous assessments of unsegmented recordings). The emotion recognition system can be more robust by weighting each frame according to their emotional saliency. From a speech production viewpoint, the approach can shed light into the underlying interplay between lexical and affective human communication across various acoustic features (Busso and Narayanan, 2007).

This study focuses on detecting emotionally salient temporal segments on the fundamental frequency. Patterson and Ladd (1172) argued that the range (i.e., the difference between the maximum and the minimum of F0 contour in a sentence or utterance) does not give information about the distribution of F0 and hence valuable emotional information is neglected. Also, according to Lieberman and Michaels (1962) low variations in F0 can be subjectively relevant in the identification of emotions. In the literature, there are some attempts to model the shape of the F0 contour. Paeschke and Sendlmeier (2000) analyzed the rising and falling movements of F0 within accents in affective speech. The study incorporated metrics related to accent peaks within a sentence. The authors found that those metrics present statistically significant differences between emotional classes. Also, Paeschke (2004) modeled the global trend of F0 in emotional speech as the gradient of linear regression. The author concluded that global trend can be useful to describe emotions such as boredom and sadness. Rotaru and Litman (2005) employed linear and quadratic regression coefficients and regression error as features to represent pitch curves. Yang and Campbell (2001) argued that concavity and convexity of the F0 contour reflect the underlying expressive state. The *Tone and Break Indices* system (ToBI) is a scheme for labeling prosody that has been widely used for transcribing intonation (Silverman et al., 1992). Liscombe et al. (2003) analyzed affective speech with acoustic features by using ToBI labels to identify the type of nuclear pitch accent, the contour type and the phrase boundaries. Despite the fact that ToBI provides an interesting approach to describe F0 contours, more precise labeling is required to generate prosodic transcripts. Taylor (2000) introduced the *Tilt Intonation Model* to represent intonation as a linear sequence of events (e.g. pitch accents or boundaries), which in turn are given by a set of parameters. However, an automatic event segmentation algorithm is required to employ this scheme and, hence, it cannot be easily applied to emotion recognition or detection tasks.

Despite current efforts to address the problem of affective speech characterization by means of modeling F0 contour, this is still an open task. The contributions of the paper concern: (a) a novel framework to detect emotional modulation based on reference templates that models F0 contours of neutral speech; (b) an insightful and thorough analysis of neutral references as a method to detect emotion in speech; (c) the generation of reference F0 contour templates with *functional data analysis* (FDA); and, (d) a study of the shortest segmentation unit that can be used in emotion detection. Extensive experiments are presented to demonstrate the discriminative power of the FDA based approach to detect emotional speech. The results on the SEMAINE database reveal that the approach captures localized emotional information conveyed in short speech segments (0.5 s). These properties of the proposed approach are interesting from the research and application points of view.

2. Emotional databases and features

2.1. Emotional databases

The analysis and results presented in Sections 3 and 4 require recordings with controlled, lexicon-dependent conditions (e.g., recordings of sentences with the same lexical content conveying different emotional states). Therefore,