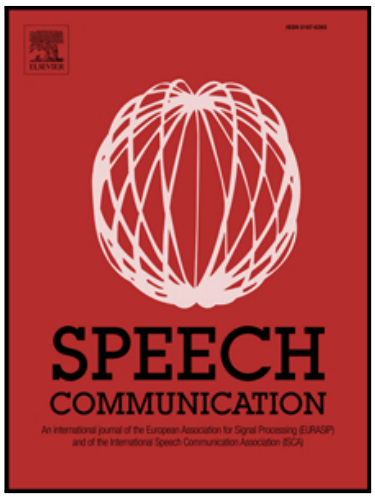

Which Prosodic Features Contribute to the Recognition of Dramatic Attitudes?

Adela Barbulescu, Rémi Ronfard, Gérard Bailly

PII: S0167-6393(16)30340-5
DOI: 10.1016/j.specom.2017.07.003
Reference: SPECOM 2474

To appear in: *Speech Communication*

Received date: 15 January 2017
Revised date: 16 July 2017
Accepted date: 28 July 2017

Please cite this article as: Adela Barbulescu, Rémi Ronfard, Gérard Bailly, Which Prosodic Features Contribute to the Recognition of Dramatic Attitudes?, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.07.003

Which Prosodic Features Contribute to the Recognition of Dramatic Attitudes?

Adela Barbulescu[1], *Rémi Ronfard*[1], *Gérard Bailly*[2]

[1]Univ. Grenoble Alpes, Inria, LJK
[2]GIPSA-lab, CNRS & Univ. Grenoble Alpes, Grenoble, France
adela.barbulescu@inria.fr, remi.ronfard@inria.fr, gerard.bailly@gipsa-lab.fr

Abstract

In this work we explore the capability of audiovisual prosodic features (such as fundamental frequency, head motion or facial expressions) to discriminate among different dramatic attitudes. We extract the audiovisual parameters from an acted corpus of attitudes and structure them as frame, syllable and sentence-level features. Using Linear Discriminant Analysis classifiers, we show that prosodic features present a higher discriminating rate at sentence-level. This finding is confirmed by the perceptual evaluation results of audio and/or visual stimuli obtained from the recorded attitudes.

Index Terms: audiovisual expressive speech; affective database; dramatic attitudes; perceptual correlates

1. Introduction

Attitudes refer to the expression of social affects and present acoustic and visual manifestations which are linked to conventions and cultural behaviors [1]. Thus, attitudes differ from basic emotional expressions, which may be seen as more spontaneous and universal expressions [2] [3].

The study of audiovisual parameters which encode the paralinguistic content of speech plays an essential role in improving the recognition and synthesis of expressive audiovisual speech. To this goal, there has been a great amount of work on the analysis and modeling of features which are found to help in the discrimination between expressive styles. Audiovisual features such as voice quality [4], acoustic prosodic features (F0, rhythm, energy) [5][6] [7], head motion [8] and facial expressions [9], have proven to be efficient in discriminating between basic emotions, attitudes or speaker identity.

While recognition of emotion, psycho-physiological state or co-verbal activities (drinking, eating, etc) is largely based on signal-based data mining and deep learning with features collected with a sliding window over multimodal frames, early studies on the expression of verbal attitudes have proposed that speakers use global prosodic patterns to convey an attitude [10][11]. These patterns are supposed to be anchored on the discourse and its linguistic structure, rather than encoded independently on parallel multimodal features. We recently evidenced the relevance of such patterns in facial displays [12].

The main aim of this work is to further explore the effectiveness of using audiovisual features at different structural levels to discriminate among expressive styles. We thus compare below the discrimination between attitudes at different structural levels (frame, syllable and sentence) and with different acoustic and visual features in order to evaluate the importance of the positioning of discriminant audiovisual events within the utterance. To that purpose, we performed a series of Linear Discriminant Analyses (LDA) on an expressive corpus of dramatic attitudes. In line with Iriondo et al [13] who used the results of a subjective test to refine an expressive dataset, we compare our best classification results with perceptual evaluation tests for the set of attitudes which are best discriminated.

The paper is structured as follows: section 2 presents approaches in related studies, section 3 presents our corpus of attitudes and the extraction of audiovisual features. Section 4 presents the experiments we carried out for automatic classification and section 5 presents the perceptual evaluation and comparison techniques, followed by conclusions in section 6.

2. Related work

Although recent years have brought a substantial progress in the field of affective computing [14], the development of emotion-modeling systems strongly depends on available affective corpora. As training and evaluation of algorithms require a great amount of data which is hard to collect, publicly available datasets represent a bottleneck for research in this field. Moreover, the majority of available datasets are limited to the six basic emotion categories proposed by Ekman [15] and include happiness, sadness, fear, anger, disgust, and/or surprise.

Databases containing affective data can be categorized under several criteria: data types used (2D or 3D visual data, speech), spontaneity (naturalistic, artificially induced or posed by professional actors or not), affective state categorization (emotion, attitudes etc). Audiovisual recording is obviously more expensive and time-consuming than audio-only recording. This is proven by the comparative amounts of publicly available audio and audiovisual datasets. For instance, the Interspeech Computational Paralinguistic Challenge [1] provides audio data from a high diversity of speakers and different languages, such as (non-native) English, Spanish, and German.

A comprehensive overview of the existing audiovisual corpora can be obtained from [24][14]. Table 1 presents a set of expressive datasets which are most relevant to our work. The works listed in the table present publicly available data that are used in several research topics: analysis, affective recognition, expressive performance generation, audiovisual conversion etc. IEMOCAP [23] and CAM3D [21] contain motion capture data of the face and upper-body posture from spontaneous performances. Large data variability is presented by Bosphorus [19] and CK+ [20] as they include more than 100 subjects posing over 20 expressions each, in the shape of action units and combinations. Another important work is the Mind Reading dataset [22] which includes video recordings of 412 expressive states classified under 24 main categories. While they serve as valuable references for the expressive taxonomy, these datasets often do not contain audio data. To our knowledge, the only publicly available affective datasets that include 3D data and

[1]http://compare.openaudio.eu/