Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 269–282
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Classification of emotion in spoken Finnish using vowel-length segments: Increasing reliability with a fusion technique

Eero Väyrynen [a,*], Juhani Toivanen [b], Tapio Seppänen [a]

[a] University of Oulu, Department of Electrical and Information Engineering, Computer Engineering Laboratory, P.O. Box 4500, FI-90014 Oulu, Finland
[b] Academy of Finland and University of Oulu, Department of Electrical and Information Engineering, Information Processing Laboratory, P.O. Box 4500, FI-90014 Oulu, Finland

Received 4 December 2009; received in revised form 16 September 2010; accepted 20 September 2010
Available online 1 October 2010

Abstract

Classification of emotional content of short Finnish emotional [a:] vowel speech samples is performed using vocal source parameter and traditional intonation contour parameter derived prosodic features. A multiple kNN classifier based decision level fusion classification architecture is proposed for multimodal speech prosody and vocal source expert fusion. The sum fusion rule and the sequential forward floating search (SFFS) algorithm are used to produce leveraged expert classifiers. Automatic classification tests in five emotional classes demonstrate that significantly higher than random level emotional content classification performance is achievable using both prosodic and vocal source features. The fusion classification approach is further shown to be capable of emotional content classification in the vowel domain approaching the performance level of the human reference.

Keywords: Automatic classification of emotion; Prosodic features; Vocal source features; Classifier fusion; Vowel segments; Spoken Finnish

1. Introduction

Automatic classification of emotional content from speech (without utilizing the lexical/semantic content) is achieving growing interest within the human–machine interaction research community. Generally, the human emotion discrimination performance is expected to be a reference, against which the automatic classification performance is measured. Important applications that the applied technological research of this kind may enable include improved human–computer interaction (e.g. the hi-tech call-center environment and automated telephone systems in general with a possibility to localize customer discontent automatically), more powerful data-mining solutions (e.g. content-based information retrieval for audio databases), and even medical applications of different kinds. A "hotspotter" that could localize potentially escalating communicative breakdowns in routine patient–caretaker interaction, making an alarm when the emotionally symptomatic speech behavior changes abruptly, would be a useful application. The successful system would monitor patient messages and prioritize them with respect to their emotional content. Entertainment applications, including robotic pets that "understand" human emotional behavior, are currently becoming increasingly available.

The automatic classification of emotional content in speech is usually based on constructing classifiers for classifying emotions in unseen emotional speech, i.e. in a speaker-independent situation. These data-driven techniques use machine learning algorithms such as neural networks or support vector machines trained on the speech prosody behavior. The used approaches have traditionally been constricted to a single classifier structure with training. The common approach has been to adopt the multiple classification task in the classification experiments; it is

* Corresponding author. Tel.: +358 (0)8 553 2526; fax: +358 (0)8 553 2612.
E-mail addresses: eero.vayrynen@ee.oulu.fi (E. Väyrynen), juhani.toivanen@ee.oulu.fi (J. Toivanen), tapio.seppanen@ee.oulu.fi (T. Seppänen).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.09.007

assumed that the emotional content of the speech data can be compartmentalized into a number of basic categories, such as angry, happy, sad, bored and tender. Recent representative studies include Dellaert et al. (1996), Batliner et al. (2000), Nwe et al. (2003), Oudeyer (2003), Ververidis et al. (2004), Morrison et al. (2007). Excellent reviews of the research are provided by Cowie et al. (2001), Bosch and L. (2003).

The major acoustic parameters which the classifiers use when distinguishing between different emotional categories in speech have been identified in a number of studies (some of which are listed above), and the number of parameters seems to vary from approximately 10 basic features (Lee et al., 2001) to about 280 (Schuller et al., 2006). In the automatic classification studies, the following acoustic/prosodic parameters have been found useful (see e.g. Banse et al., 1996; Scherer, 2003; Morrison et al., 2007): f0-related features (e.g. mean, minimum, maximum, quartiles, difference of quartiles, average f0 change during fall/rise), speaking rate related features (e.g. average duration of voiced segments, average duration of unvoiced segments, average duration of quartiles, difference of quartiles), and spectral features (e.g. amount of energy at different frequencies). By adding new mathematical derivatives of these parameters, it is easy to come up with a very high absolute number of features.

Recently, a fusion of multiple speech signal derived classifiers in a multimodal classification approach has gained some attention. This fusion approach was first introduced in the field of speaker identification/verification where it was proposed by Campbell et al. (2003). Typically the acoustic/prosodic features used in the fusion have been a some combination of mel frequency cepstrum coefficients (MFCC), pitch contour features, phonetic features, and lexical features. The choice of features has been directly adapted to the field of emotion recognition. A Spanish emotion classification study by Barra et al. (2006) includes a fusion scheme using MFCC and prosodic features. Another study for English speech using similar fusion approach and features was published by Kim et al. (2007).

In emotion classification/discrimination studies, both computer- and human-based, the speech context has usually been restricted to either a sentence or a short passage. Very typically, utterances lasting 1–10 s are used as basic emotion-carrying units in the experiments. Some relevant datasets are listed below. The Kismet speech corpus (Breazeal and Aryananda, 2002) contains (acted) American English utterances of variable length (1.8–3.3 s), with emotions such as approval, attention, prohibition, soothing and neutral. The BabyEars speech corpus (Slaney and McRoberts, 2003) contains American English natural parent–infant interaction (parents addressing their children), with content classes such as approval, attention and prohibition. The length of the utterance is 1–9 s in the BabyEars dataset. The MediaTeam Emotional Speech Corpus (Seppänen et al., 2003a) consists of acted utterances in standard Finnish, with an average duration of approximately 10 s (with neutral, angry, happy and sad emotional content). Other studies using (relatively short) sentences as a basis for the automatic classification experiments include Paeschke and Sendelmeier (2000), with simulated neutral, angry, fearful, joyful, sad, disgusted and bored utterances in German; Engberg and Hanse (1996), with simulated neutral, surprised, happy, sad and angry utterances in Danish; McGilloway et al. (2000), with simulated angry, happy, fearful, sad and neutral utterances in English; and Polzin and Waibel (2000), with sad, angry, and neutral utterances from English movies. It seems that, in general, the speech data used in experiments on the automatic classification of emotions represents spoken language at the clause/phrase/utterance level, i.e. typically involving a speech unit with a coherent intonation pattern. For instance, the Kismet speech corpus largely consists of utterances (representing exhortations, vocatives, etc.) with a coherent (discoursally determined) prosodic structure although the speech stimuli, as such, can be very short in duration (1–2 s).

We assume that an emotionally laden utterance functioning as an independent information carrying phrase can indeed be very short, a vowel-level speech event, also in spoken Finnish. Hence, such a unit would be carrying the typical prosodic/intonational characteristics of a longer stretch of speech, in addition to the voice quality features (it should be mentioned that the role of intonation contour prosodic features in emotion signaling in connected spoken Finnish are understood relatively well (see Toivanen et al., 2004). In the existing literature on the vocal characteristics of emotion at the vowel level in Finnish, the main focus has so far been on the vocal source parameters (Airas et al., 2005; Laukkanen et al., 1996). We assume that, under certain circumstances – for example functioning as a brief command, an emotional outburst in lieu of a phrase or a unit of back-channel communication – a vowel-level speech event can attract phrase-level intonation contour parameters. As is well-known, intonational features can convey highly elaborate emotional/attitudinal meanings in spoken language (Cruttenden, 1997; Mozziconacci and Hermes, 1999). The intonation contour features used in the classification experiments in this paper (see Table 2 below) can be seen as reflecting the dynamicity of intonation variation in speech; for example, Cruttenden's (1997) terms "accent range", "complexity", "key" and "register" obviously reflect the dynamicity aspects of intonation contours.

From the viewpoint of spoken Finnish, the stereotypical view has been that in utterances there is very little intonation, at least in non-emotional discourse. A fortiori, it has been assumed that all prosodic phenomena at the syllable level can be accounted for in terms of word stress only. For instance, Vroomen et al. (1998) argue that, as Finnish has fixed word-initial word stress, with a fixed f0 rise, intonational features per se play a somewhat minor role at the syllable level although the authors do acknowledge that "it seems possible that a stressed syllable can be perceived as stressed without reference to neighboring syllables, for example on the basis of characteristic f0 transitions within