Contents lists available at ScienceDirect
Digital Signal Processing
www.elsevier.com/locate/dsp

Prominence features: Effective emotional features for speech emotion recognition

Shaoling Jing, Xia Mao, Lijiang Chen *

School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Prominence features
Speech annotation
Consistency assessment
Speech emotion recognition

ABSTRACT

Emotion-related feature extraction is a challenging task in speech emotion recognition. Due to the lack of discriminative acoustic features, classical approaches based on traditional acoustic features could not provide satisfactory performances. This research proposes a novel type of feature related to prominence, which, together with traditional acoustic features, are used to classify seven typical different emotional states. To this end, the author group produces a Chinese Dual-mode Emotional Speech Database (CDESD), which contains additional prominence and paralinguistic annotation information. Then, a consistency assessment algorithm is presented to validate the reliability of the annotation information of this database. The results show that the annotation consistency on prominence reaches more than 60% on average. Subsequently, this research analyzes the correlation of the prominence features with emotional states using a curve fitting method. Prominence is found to be closely related to emotion states, to retain emotional information at the word level to the greatest possible extent and to play an important role in emotional expression. Finally, the proposed prominence features are validated on CDESD through speaker-dependent and speaker-independent experiments with four commonly used classifiers. The results show that the average recognition rate achieved using the combined features is improved by 6% in speaker-dependent experiments and by 6.2% in speaker-independent experiments compared with that achieved using only acoustic features.

1. Introduction

In recent years, technology for automatic emotion recognition from speech has matured sufficiently to be applied in several real-life scenarios [1–3], such as call centers [4,5], auxiliary disease diagnosis [6–9], remote education [10–12], driving safety [13,14], and computer games [15]. However, existing speech emotion recognition (SER) technology has not achieved very good performance, probably because of the lack of effective emotion-related features. To solve this problem, we explore the effect of prominence features in SER.

There is a diversity of viewpoints regarding the definition of prominence. Phonetician Zhao Li holds that prominence, simply speaking, refers to the fact that some syllables sound more prominent than the surrounding syllables in a speech stream, which is caused by different intensities that the speaker put on different syllables [16]. Streefkerk et al. regard prominence as the perceptual salience of a language unit. Sentence accent and pitch accent lead to perceived prominence, and there is no definite boundary between sentence accent and pitch accent [17]. Meanwhile, Terken defines prominence as "words or syllables that are perceived as standing out from their environment" [18]. Kakouros et al. also agree with this definition [19]. In Chinese, a word is defined as a syllable. Traditionally, Chinese prominence has lexical stress and sentence accent. Hence, we use the term prominence here to refer to the accentuation of syllable within words or of syllables within sentences.

Speech prominence, which refers to a prosodic property, is typically viewed within the contexts of phonetic properties and physical realization [20]. In some languages, prominence can be a very subjective topic. If we observe the prominence trajectories of only a few speakers, it is likely that these prominence trajectories will vary greatly from one another. However, statistical methods are needed to study the subjectivity of prominence. In particular, the general idea is that the subjective perception of prominence is a response elicited to unpredictable prosodic trajectories in a normal train of speech, which draw the attention of the listener [19]. Indeed, Kakouros et al. have proven that the statistical distribution of prosodic cues can impact the subjective perception of word prominence [19]. Therefore, the subjectivity of prominence makes this research a challenging work. However, prominence has been

* Corresponding author.
E-mail address: chenlijiang@buaa.edu.cn (L. Chen).

https://doi.org/10.1016/j.dsp.2017.10.016
1051-2004/

proven to be useful in many spoken language applications since it encompasses affective expression and affects speech understanding, especially in speech synthesis [21,22]. Researchers have proven that integrating prominence patterns into speech recognition can improve speech classifier performance [23,24]. However, so far, few studies have applied prominence attributes to improve the accuracy of emotion recognition. Inspired by the use of prominence by speakers to express their emotions or attitudes, we explore whether prominence has an impact on SER. For this purpose, we need to extract the prominence features from utterances. However, the majority of studies on automatic prominence detection have focused on supervised methods, which require manually annotated prominence labels [25]. Therefore, the objective of our work is to create prominence annotations, extract prominence features and apply them in speech emotion recognition.

1.1. Related work

Generally, valid speech corpora are usually accompanied by emotional labels, which are used to train models for SER. Most of existing Chinese emotional speech databases have three main problems. First, they are not available/public, thus not accessible by researchers. Second, they are annotated for specific tasks, thus do not contain enough information for general speech emotion recognition task. Third, their annotations are confined to emotional labels and transcription information, while corpora of other languages contain much more rich annotation information. For instance, the German database emoDB is annotated with phonetic transcriptions, stress, and segment and pause boundaries [26]. The Japanese database UU is annotated with orthographic transcriptions, dimensional emotions (such as pleasant–unpleasant and aroused–sleepy) and paralinguistic information [27]. However, so far, the available Chinese emotional speech databases lack such rich annotation information. For example, Chinese emotional speech database DFEIC is only annotated with prosody, function and emotional states [28]. Paralinguistic information refers to meaningful information, such as emotion or attitude, delivered along with linguistic messages. Therefore, to study whether the combination of paralinguistic information, prominence and sentence function is useful for SER, the author group produces a new Chinese Dual Modal Emotional Speech Database (CDESD), which contains rich annotation information including text transcription, syllable segment, initials, finals, tones, silence, voiced and unvoiced segments, prominence, sentence function and emotional states. In our emotional annotation, an utterance may express more than one emotion, in that case the emotional state label will contain all emotion states, unlike in the other databases.

Since follow-up experiments are based on the assumption that the annotation information is correct, we must ensure that the labeled data is reliable before using them in SER. In light of the appraisal theory from the domain of emotional psychology [29], each annotator may have his or her own subjective perception of the affective state expressed by an individual, thus the annotation may be influenced by the identity of the annotator, annotator's experience, memories, and reasoning, etc. [30,31]. Besides, the annotation task is very time-consuming and laborious; hence, the long time required for annotation may affect the subjective perceptions and judgments of the annotators. To ensure the reliability of annotated data, we need to use relevant algorithms to determine the agreement among the annotators. Cohen's kappa [32] is a measure of the agreement between two raters when determining to which categories a finite number of subjects belong. The two raters either agree in their rating (i.e., the category to which a subject is assigned) or they disagree. However, Cohen's kappa does not account for the degree of disagreement. The weighted Cohen's kappa measure [33] addresses this issue by using a predefined table of weights that measure the degree of disagreement between the two raters. However, neither method is suitable for our annotation scheme. For the situation at hand, one of rater's jobs is to determine the boundaries between different syllables. Since each boundary's position can be selected from a continuous region, it does not fall into a fixed number of classes. Besides, in some annotation layer, such as syllable and tones layer and initials, finals and tones layer (see Section 3.1), each syllable may be divided into uncertain number of classes and the division position is also uncertain. Therefore, it is necessary to present a consistency assessment algorithm specific to our situation.

After confirming the reliability of the prominence annotation, we explore whether prominence can be used as a basis for extracting useful features to improve SER performance. Most previous studies have relied on acoustic features, such as pitch, energy and spectral coefficients [34–38]. However, in recent years, many researchers have begun to incorporate additional information to improve the accuracy of emotion recognition. Lee et al. combined acoustic and language information for emotion recognition, thereby improving negative emotion classification by 45.7% and 32.9% compared with the use of only acoustic information and only language information, respectively [39]. Schuller et al. proposed to perform SER by combining acoustic features with linguistic information [13]. Chen et al. proposed two classes of speech emotional features extracted from electroglottography and speech signals [40], and combined those features with traditional features for SER [41]. Mencattini et al. proposed novel features related to the amplitude modulation in a speech emotion prediction model [42]. Moreover, paralinguistic and non-linguistic information [27, 43,44], emotional point information [45], and lexical and semantic information [46] are also closely related to the emotional states of human beings and play a crucial role in SER. In particular, prominence is an important property of speech. Speakers may use prominence to draw the listener's attention to specific points of an utterance to express their emotions or attitudes. Previous works have demonstrated the usefulness of prominence in clarifying ambiguities in spoken utterances and constructing understandable synthetic speech [47]. Therefore, it is feasible to use prominence in SER system.

1.2. Main contributions

This research aims at producing a new speech emotion recognition system based on traditional acoustic features and our newly proposed prominence features. To this end, we identify three aspects of the problem related to prominence features in SER: (1) prominence annotation, (2) annotation consistency, and (3) prominence effectiveness.

Fig. 1 presents the scheme of the proposed recognition system. First, each syllable is annotated with its prominence, and a consistency assessment algorithm is designed to validate the reliability of the annotation. Second, the manually labeled prominence information is extracted, the extracted prominence information is pre-processed to obtain the prominence features, and a curve fitting-based (CF-based) method is used to analyze the correlation between the proposed features and emotional states. Third, 290 acoustic features are extracted, and the most useful features are selected. The Sequence Backward Selection (SBS) algorithm and the Sequence Forward Selection (SFS) algorithm are used for feature selection. Fourth, the prominence features are combined with acoustic features for use in classifying emotional states. The Double Layer Feature Dimension Reduction (DLFDR) model, which combines Principal Component Analysis (PCA) and Nonparametric Discriminant Analysis (NDA) [48], is applied to reduce the feature dimensions. Finally, the novel features are tested in speaker-independent and speaker-dependent experiments. Good perfor-

Please cite this article in press as: S. Jing et al., Prominence features: Effective emotional features for speech emotion recognition, Digit. Signal Process. (2017), https://doi.org/10.1016/j.dsp.2017.10.016