Speech Emotion Recognition Using Spiking Neural Networks

Cosimo A. Buscicchio, Przemysław Górecki, and Laura Caponetti

Universita degli Studi di Bari, Dipartimento di Informatica
Via E. Orabona 4, 70126, Bari, Italy
{buscicchio, przemyslaw, laura}@di.uniba.it

Abstract. Human social communication depends largely on exchanges of non-verbal signals, including non-lexical expression of emotions in speech. In this work, we propose a biologically plausible methodology for the problem of emotion recognition, based on the extraction of vowel information from an input speech signal and on the classification of extracted information by a spiking neural network. Initially, a speech signal is segmented into vowel parts which are represented with a set of salient features, related to the Mel-frequency cesptrum. Different emotion classes are then recognized by a spiking neural network and classified into five different emotion classes.

1 Introduction

Human social communication depends largely on exchanges of non-verbal signals, including non-lexical expression of emotions in speech. Speech supplies a reach source of information about a speaker's emotional state and the research in the area of emotion recognition is important for developing emotion based human-computer interactions [1]. In fact, recent experiments [2] show that humans use the schemes of interpersonal interaction also when they interact with their computers.

Some early physiological experiments made by Petrushin [3] have shown that humans are able to classify emotions correctly with an average accuracy of 65%. Performance of the artificial emotion recognition systems largely depends on the extraction of relevant features from speech. In the field of signal analysis, many traditional and recent works are based on the analysis of "prosodic" information, which includes pitch, duration and intensity of utterance [4]. For example, Chiu [5] extracted a number of features from speech and a multilayered neural network was employed for the classification. Dellaert [6] compared different classification algorithms and feature selection methods. They achieved 79.5% accuracy with four categories of emotions and five speakers uttering 50 short sentences per category.

In this work, we propose a biologically plausible methodology for the problem of emotion recognition, based on the extraction of vowel information from an input speech signal and on a spiking neural network classifier. Since the parts of

F. Esposito et al. (Eds.): ISMIS 2006, LNAI 4203, pp. 38–46, 2006.

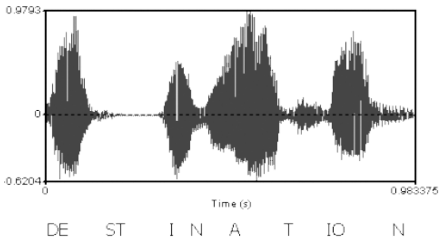

Fig. 1. Sound waveform of the word "destination" for an "angry" stress condition utterance from SUSAS dataset

the speech signal containing consonants do not carry information about emotion, vowel parts are extracted from the speech samples and are represented with a set of salient features, related to the Mel-frequency cesptrum. Different emotion classes are recognized by a spiking neural classifier working on the previously extracted features. The effectiveness of our approach is demonstrated during the experimental session performed with samples selected from the SUSAS database.

The paper is organized as following. In section 2 the feature extraction process is presented. Section 3 describes the details of the spiking neural network classifier. Section 4 presents the network training and classification process. Sections 5 and 6 contain details of the experimental session and the conclusion.

2 Feature Extraction

This section presents our strategy for feature extraction from sentences of spoken words. It is based on the assumption that the sentences are already segmented into separate words. Since the consonants of the word do not carry information about emotions [3], we consider only the vowels. In order to extract vowels from a sentence, Brandt's Generalized Likelihood Ratio (GLR) method is used [7], based on the detection of signal discontinuities. In this way, each sentence is represented as a set of N vowel segments. As an example, the waveform of a word "destination" is shown in figure 1. Note that the vowels correspond to the louder regions of the waveform.

Successively, each segmented vowel is analyzed in order to extract a set of salient features that preserves most of the information. In this work we have considered the Mel-frequency cepstral coefficients [8] as a representation of the vowel. Cepstrum analysis measures the periodicity of the frequency spectrum of a signal that provides information about rate changes in different spectrum bands. The Mel-frequency cepstral coefficients represent cepstrum information in frequency bands, positioned logarithmically on the Mel scale. This scale corresponds to the particular range of pitches that were judged by listeners to be equidistant one from the other. The Mel scale is given by the following formula: