Speech Emotion Recognition: Features, Techniques, and Evaluation

Maria Art Antonette D. Clariño
Institute of Computer Science
College of Arts and Sciences
University of the Philippines Los Baños
mdclarino@up.edu.ph

ABSTRACT

This paper presents commonly used techniques, features used as emotion descriptors, and performance evaluation in speech emotion recognition. Auditory data poses a quantitative manner of representing emotion. Characteristics of data (collected from specific type of subjects, natural or acted) should match the goal of a speech emotion recognition system. Established techniques in machine learning have been employed to model and classify emotions such as support vector machines, neural networks, gaussian mixture models, hidden Markov models, and k-nearest neighbors. Under prosodic feature type, pitch and intensity are commonly used while Mel-Frequency Cepstral Coefficients well represent spectral features. Evaluation of classifier performance must also be focused on as it would justify the selected combination of technique, features, and emotions in the system. Classification tasks can be classified depending on the number of classes and type of classification. There are four categories namely binary, multi-class, multi-labelled, and hierarchical [55]. All of these components, when combined effectively, constitute to a successful speech emotion recognition system.

CCS CONCEPTS

•General and reference →Surveys and overviews; •Computing methodologies →Natural language processing; Speech recognition;

KEYWORDS

Speech Emotion Recognition, Auditory Models, Survey

1 INTRODUCTION

With technological advances, much efforts are focused on the objective of enabling machines to function as close to human beings do. From computer vision emulating how human beings see to robotics mimicking how human limbs move, many studies have been made to achieve this objective. The most common and natural way of communication by humans is through speech or spoken language [42][64][3].

Auditory data poses a quantitative manner of representing emotion. This property enables matching of quantitative descriptors to specific emotions towards the classification process. Machine learning techniques have been exhaustively employed in natural language processing problems. Data may be in text or spoken language. In the case of speech, it entails a specific class of features to model a class. It begins with understanding that speech information can provide linguistic (explicit information such the language code and semantics) and paralinguistic information where the emotions conveyed in the speech falls under [48] [40].

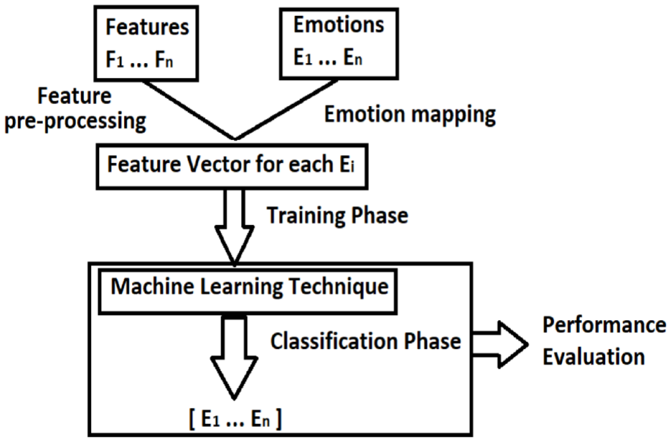

Figure 1: Diagram of emotion recognition system

Figure 1 illustrates the basic elements and stages of a supervised emotion recognition system. It starts with the selection of features as descriptors and emotions as classification categories. Most commonly used features are pitch and energy contours [42]. A feature vector is created for each emotion as a descriptor that would be used during the classification phase. In a supervised system, the same feature vector is used for the training phase wherein these features are given values mapped to a particular emotion. During the classification phase, if the test data matches the values obtained during training phase by some measure then the test data is labeled to be portraying a certain emotion. Given a test data, it will be classified depending on the selected technique. Most common techniques used are support vector machines (SVM), hidden Markov models (HMM), neural networks (NN), and k-nearest neighbour (KNN) algorithm. These techniques as well as multiple layers perception have been regarded in numerous studies to provide good results in different fields of research [43]. These methods may be explored individually or combined evaluating their classification prowess within the context of speech emotion recognition (SER), a multi-class problem. This notion is supported by [43] claiming that no specific method is set as best used for SER, thus, encourages much exploration.