EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Biologically inspired emotion recognition from speech

Laura Caponetti[1*], Cosimo Alessandro Buscicchio[2] and Giovanna Castellano[1]

Abstract

Emotion recognition has become a fundamental task in human-computer interaction systems. In this article, we propose an emotion recognition approach based on biologically inspired methods. Specifically, emotion classification is performed using a long short-term memory (LSTM) recurrent neural network which is able to recognize long-range dependencies between successive temporal patterns. We propose to represent data using features derived from two different models: mel-frequency cepstral coefficients (MFCC) and the Lyon cochlear model. In the experimental phase, results obtained from the LSTM network and the two different feature sets are compared, showing that features derived from the Lyon cochlear model give better recognition results in comparison with those obtained with the traditional MFCC representation.

1. Introduction

For many years human-computer interaction researchers focused their attention on the synthesis of audio-visual emotional expressions. Despite the great progress made in human-computer interaction, most of the existing interfaces are unidirectional and unfriendly in the sense that they do not allow the computer to understand the emotional state of the user [1-3]. During the recent years, many researchers have faced the problem of designing interfaces that are able to express and also to perceive emotions. This has introduced a specific research field, known as Speech Emotion Recognition, which is aimed to extract the emotional state of a speaker from his or her speech.

Speech emotion recognition is useful for applications requiring natural man-machine interaction, such as web movies and computer tutorial applications, where the response to the user depends on the detected emotions.

In order to perform recognition of speech emotion, two issues are of fundamental importance: the role of speech features on the classification performance, and the classification system employed for recognition [4].

Since humans have far superior recognition abilities than any other existing artificial classification system, a common research trend is to simulate the recognition mechanisms of biological systems to design artificial systems having a greater degree of fidelity with respect to the biological counterparts. This design approach involves the study of the physiology of biological systems and the research of methods that are able to model the particular biological structures.

Artificial neural networks are based on this idea, since they were designed to mimic the biological neural networks found in the human brain. They are formed of groups of artificial neurons connected together in much the same way as the brains neurons. Connections between artificial neurons are determined by learning processes, similarly as connections between human brain neurons are determined and modified by learning and experience acquired during time. Commonly, feed-forward neural networks are employed, in which neurons are connected using a "feed-forward" structure that allows signals to travel from input to output only. Recurrent neural networks, including loop connections that allow signals to travel both directions, are potentially very powerful and more biologically plausible than feed-forward neural networks. This encourages the application of recurrent neural networks to complex recognition tasks.

Starting from this idea, several emotion recognition approaches have been proposed in the literature (see Section 2), which use a specific recurrent neural network called long short-term memory (LSTM) [5], whose model is closely related to a biological model of memory in the prefrontal cortex.

* Correspondence: laura@di.uniba.it
[1]Dipartimento di Informatica, Università degli Studi di Bari, Via E. Orabona 4, 70126, Bari, Italy
Full list of author information is available at the end of the article

In this article, we present an approach for emotion recognition from speech that combines LSTM architecture with two different representation models of the emotion speech signal: the Lyon cochleagram model, and the more classic Mel frequency cesptral representation. These two representations are compared with a view to unveil differences between the two models in terms of emotion recognition rate.

The basic idea of the proposed approach is to use biologically inspired models not only for emotion recognition but also for signal representation. This is in line with recent study in the literature, which faced the problem to model and recognize the sound on the basis of the human ear and human brain studies [6]. Specifically, our idea is to use a classifier, modeled on the human brain combined with a signal representation modeled on the human auditory system. In the proposed approach, two different biologically inspired representations are investigated and compared. The first is based on the traditional mel-frequency cepstral coefficients (MFCC) [7], which are extracted by filters whose bandwidths are distributed with respect to a nonlinear frequency scale according to the human perception of pitch. The other representation considered in our study is the Lyon Cochlear model [8], which is based on the study of the human auditory system and models the behavior of the cochlea, or inner ear.

The rest of the article is organized as follows. Section 2 presents a brief overview of the research in the field of speech emotion recognition. The description of the proposed approach is provided in Section 3. The experimental results and discussion are provided in Section 4. Finally, some conclusions are provided in Section 5.

2. Background

Many theories of emotion have been proposed [9], and some of these have not been verified until some measurements of physiological signals have become available. In general, emotions are short-term states, whereas moods are long-term states, and temperaments or personalities are very long-term states [10].

Emotional states are often correlated with particular physiological states [11], which in turn present predictable effects on speech features, especially on pitch, timing, and voice quality. For instance, when one is in a state of anger, fear, or joy, the sympathetic nervous system is aroused, the heart rate and blood pressure increase, the mouth becomes dry, and there are occasional muscle tremors. Speech is then loud, fast and enunciated with strong high-frequency energy. When someone is bored or sad, the parasympathetic nervous system is aroused, the heart rate and blood pressure decrease, and salivation increases, which results in slow, low-pitched speech with a weak high-frequency energy. In [12], the authors have shown that physiological effects are rather universal, that is, there are common tendencies in the correlation between some acoustical features and basic emotions across different cultures. For instance, Tickle [13] performed some experiments to show that there is little performance difference between detecting emotions expressed by people speaking the same language or different languages. These results indicate that emotion recognition from speech could be performed independently of the language semantics. For this reason, in this study, we are not interested in recognizing words. Rather, we want to recognize the emotions expressed during the pronunciation of words, namely, we consider the recognition of emotions in speech signals.

In order to define an artificial system capable of recognizing emotions from speech, three main issues have to be addressed: how to determine the emotions to be recognized, how to represent them, and how to classify them. Each issue can be addressed by different approaches. A brief overview is given in the following.

2.1. Emotion identification

This problem can be seen as an emotion labeling problem, which requires different emotions to be clustered into few emotion categories. A discussion of the literature describing human vocal emotion, and its principal findings, is presented in [14,15]. Usually, two different methods are used to label emotions. The first approach is to associate emotions to labels denoting discrete categories, i.e., human judges choose from a prescribed list of word labels, such as anger, disgust, fear, joy, sadness, and surprise. One problem with this approach is that speech signals may contain blended emotions. In addition, the choice of words may be too restrictive, or culture dependent. The other approach is to consider multiple dimensions or scales to describe emotions. Instead of choosing discrete labels, observers can indicate their impression of each stimulus using several continuous scales, for example, pleasant-unpleasant, attention-rejection, simple-complicated, etc. Two common scales are valence and arousal. Valence describes the pleasantness of the stimuli, with positive (or pleasant), on one hand, and negative (or unpleasant) on the other. For example, happiness has a positive valence, while disgust has a negative valence. The other dimension is arousal or activation. For example, sadness has a low arousal, whereas surprise has a high arousal level. The different emotional labels could be plotted at various positions on a two-dimensional (2D) plane spanned by these two axes to construct a 2D emotion model [16]. Scholsberg [17] suggested a 3D model in which the attention-rejection dimension is added to the valence and arousal dimensions.

2.2. Emotion representation

Another important problem in emotion recognition from speech is the extraction of significant features

from the speech signal to define salient characteristics suitable to perform emotion classification. The problem to explore which features could describe better emotions has been widely addressed [18-20]. All authors agree that the most crucial aspects of emotion are related to *prosody*. Prosody can be considered a parallel channel for communication, carrying some information that cannot be simply deduced from the lexical channel. Prosody is related to pitch contour (in this context called F0), intensity contour, and duration of utterances [21]. Therefore, muscle motions transmit all aspects of prosody. Main factors that influence the pitch are the vocal fold tension and the subglottal pressure. These are both smoothly changing functions of time, controlled by nerve impulses, Newtonian mechanics, and the viscoelasticity of tissue [22]. The overall relationship between muscle activation and pitch is smooth, nearly linear, and the effects of the different muscles can be combined into smooth frequency changes. Detailed physiological models for F0 are described in [23]. Although pitch and energy are the most important features describing prosody [24,25], duration and amplitude are also important prosody features, as indicated in [26,27]. Others methods to derive features representing emotions analyze the signal in frequency bands defined with different auditory models. In this study, we adopt two different representations of the signal, one is based on the Lyon cochleagram model. and the other on the more classic mel frequency cesptral representation.

2.3. Emotion classification

Given a set of features describing the emotion in speech, the other important issue in the development an automatic speech emotion recognizer is the choice of the classification method. Various types of classification models have been used for this task, such as the hidden markov model (HMM), the Gaussian Mixture Model, the neural networks, and the support vector machines (SVM) [11,28,29].

Dellaert et al. [30] compare different classification algorithms and feature selection methods. They achieve 79.5% accuracy with four emotion categories and five speakers speaking 50 short sentences per category. In [31], some tests of human performance in recognizing emotions in speech are performed, obtaining an average classification rate of about 65%.

Chen [32] proposes a rule-based method for classification of audio data into one of the following emotion categories: happiness, sadness, fear, anger, surprise, and dislike. The audio data were derived from speech of two speakers: one speaking Spanish, and the other speaking Sinhalese. These languages were chosen to avoid subjective judgments to be influenced by the linguistic content, as the listeners did not comprehend either of the two languages.

In [33], those authors obtained preliminary results using MFCC representation and spiking neurons network that recognizes short, complex temporal patterns with a recognition percentage of 67.92%.

In [34], those authors reported results obtained using HMM under the NATO project "speech under stress," who aimed to obtain reliable stress measures and to study the effect of speech under stress on the performance of speech technology equipment. A variety of calibrated data are collected under realistic uncontrolled conditions or simulated conditions. Data under simulated conditions are collected in two databases, the SUSAS and DLP databases, which include simulated stress by asking subjects to respond to an externally controlled condition, such as speaking rate (DLP), or speaking style (SUSAS), or dual-tracking computer workload (SUSAS). Parameters indicating a change in speech characteristics as a function of the stress condition (e.g., pitch, intensity, duration, and spectral envelope) are applied to several samples of stressed speech. The effect on speech obtained for perceptual noise and some physical stressors is evaluated. It is shown that the effect of stressed speech on the performance of automatic speech recognizers and automatic speaker recognizers is marginal for some types of stress (DLP), while the speaking style has a major effect. In the speaker recognition task, their evaluation shows that when stress is present, the recognition rates decrease significantly especially for speech under loud and angry condition.

In the recent years, the LSTM neural networks [35] have become a new effective approach to support application of speech analysis and recognition in which the modeling of long time dependencies is relevant. LSTM neural networks has been proven to efficiently learn many difficult tasks involving recognition of temporally extended patterns in noisy input sequences and extraction of information conveyed by the temporal distance between events.

Several approaches using LSTM networks have been proposed. In [36], Graves has proposed to apply a long short-term memory (LSTM) architecture to speech recognition to provide a more robust and biologically alternative to statistical learning methods such as HMMs. The reported results are comparable to the HMM-based recognizers on both RIDIGTS and TI46 speech corpora.

In [37], an approach for continuous emotion recognition based on LSTM network is introduced, where emotion is represented by continuous values on multiple attribute axes, such as valence, activation, or dominance. This approach includes modeling of long-range dependencies

between observations, and thus outperforms techniques like support-vector regression. In their study [37], those authors used the HUMAINE database, containing data extracted from audio-visual recordings of natural human-computer conversations and labeled continuously in real time by four annotators with respect to valence and activation. The same authors provide results with different classifiers and show that classification by LSTM network is as good as human annotation, which confirms that modeling long-range time dependencies is advantageous for continuous emotion recognition.

Using the same emotional space as in [37], the authors of the study [38] investigate a data-driven clustering approach based on k-means to find classes that better match the training data and to model the effective states that actually occur in the specific recognition task. Firstly, a number of emotional states are identified by clustering, and then an LSTM network is used to recognize an emotional state between those predetermined by clustering. The results of the latter show that discriminative LSTM outperforms standard SVM. Finally, in another study [39], a bidirectional LSTM is successfully used for emotion recognition related tasks, such as keyword spotting in emotionally colored spontaneous speech.

As an extension of the above last approaches, in this article, we propose to apply LSTM architecture to two different representation models of emotion speech signal: the Lyon cochleagram model, and the more classic mel frequency cesptral representation. These two representations are compared, with a view to unveil differences between the two models in terms of emotion recognition rate.

3. Biologically inspired approach

The main aim of this study is to define an approach for the emotion recognition from speech taking into account biologically inspired methods for signal representation and classification. The general framework of our approach, depicted in Figure 1, consists of three main phases: preprocessing, representation, and emotion classification.

Since we intend to recognize emotions expressed during the pronunciation of spoken words, we consider only the vowel and the voiced consonant parts of the words. In fact, the parts containing unvoiced consonants do not carry any information about emotions[1], thus significantly reducing the complexity of the input signals. Then, in order to extract the voiced parts of the word, each utterance is segmented using the Brandt's generalized likelihood ratio (GLR) method, based on detection of discontinuities in a signal [40]. More specifically, the preprocessing phase of the proposed approach is to divide the speech signal into small segments. Each segment can be further divided into a number of time frames in which the signal is considered to be approximately stationary. Then, each frame can be described using spectral features as a short-time representation for the signal. It is recognized that the emotional content of an utterance has an impact on the distribution of the spectral energy across the speech range of frequencies.

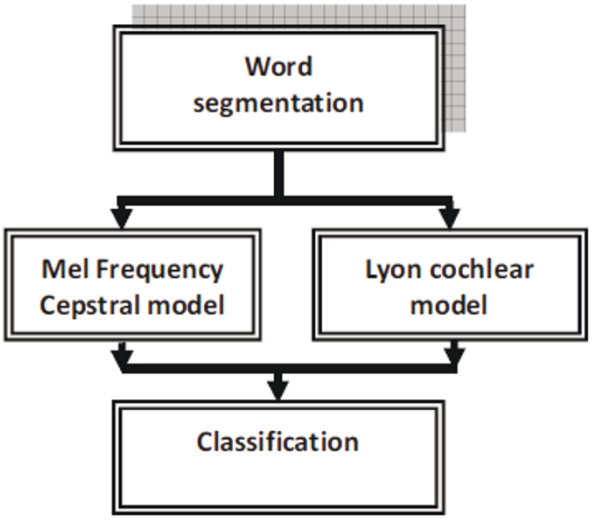

Figure 1 Functional diagram of the emotion recognition system.

In this study, features for signal representation are derived from two different auditory models: the mel-frequency cepstral model, and the Lyon cochlear model. Both models are biologically inspired. The first is based on Cepstrum analysis that measures the periodicity of the frequency spectrum of a signal, which provides information about rate changes in different spectrum bands. The MFCC represent cepstrum information in frequency bands positioned logarithmically on the mel frequency scale [7], which is a particular range of pitches judged by listeners to be equal in distance from one another. The MFCCs are based partly on an understanding of auditory perception: the log energy scale matches the logarithmic loudness perception of the ear, and the mel frequency scale is based on pitch perception. The mel frequency cepstral representation is well known in the literature, thus no further detail is given here. The second considered model, namely the Lyon cochlear model, is described in detail hereafter. Also, the LSTM recurrent network used for emotion classification is described henceforth.

3.1. Lyon cochlear model

Lyon and Mead [8] presented a multi-level sound analysis algorithm which models the behavior of the cochlea, or inner ear, in much better detail than any other sound

analysis or speech analysis algorithms. The model can also be viewed as an approach to speech analysis based on the physiology of hearing, as opposed to more popular approaches based on the physiology of speech.

More specifically, Lyon found that active processes in the cochlea could be modeled by tuning each section to have a small resonant frequency band, in which the gain from input to output is slightly larger than unity. Such a model is not sharply tuned; no single filter stage has a highly resonant response. Instead, a high gain effect is achieved by the cumulative effect of many low gain stages.

The Lyon model combines a series of notch filters with a series of resonance filters, which, at each point in the cochlea, filter the acoustic wave. Each notch filter operates at lower frequencies; therefore, the net effect is to gradually low-pass filter the acoustic energy. An additional resonator (or band-pass filter) picks out a small range of the traveling energy and models the conversion into basilar membrane motion (Figure 2). This motion of the basilar membrane is detected by the inner hair cells. A combination of a notch and a resonator is called a stage (Figure 3). The output of the Lyon cochlear model, named cochleagram, is a matrix of floating point values, in which every column is a time frame and every row is a frequency band. Each value (i, j) of this matrix provides the energy of the ith frequency band as a function of the jth time frame.

3.3. LSTM neural network

Using a neural network to process a time-dependent signal, such as speech, requires splitting the signal into time windows and treating the inputs as spatial [41]. Application of time-windowed networks to speech recognition tasks introduces two major problems. First, a fixed dimension for the time window has to be determined. Large windows lead to a high number of network inputs, with consequent long training time and high network complexity. Conversely, small windows may ignore long time dependencies such as the position

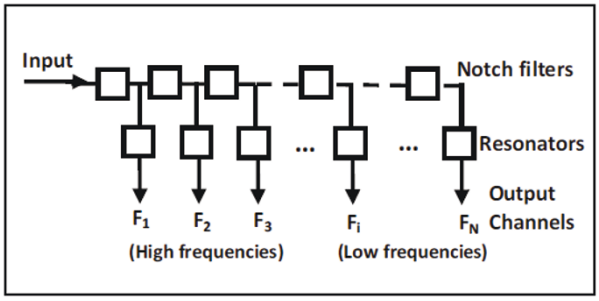

Figure 2 Block diagram of the filterbank used in the Lyon cochlear model.

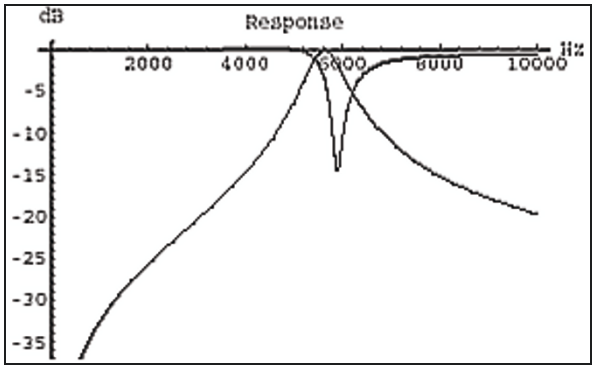

Figure 3 Responses of notch and resonance filters, combined to form a stage.

of a word in a sentence. Second, often time-windowed networks are inflexible with regard to temporal changes in the rate of input (nonlinear time warping).

These problems can be overcome through the use of recurrent neural networks that do not transform temporal patterns into spatial ones. Rather, recurrent networks process a time-dependent signal one frame at a time, using feedback connections that create a memory of previous inputs.

This is in analogy to biological neural networks, which are highly recurrent (although some functions, in particular sensory channels, are thought to be effectively feed-forward). The human brain can be modeled as a recurrent neural network, which is a network of neurons with feedback connections. Therefore, recurrent neural networks represent a valid biologically inspired approach to speech recognition that overcomes such problems as long time dependencies and temporal distortion.

In this article, we investigate the use of a particular recurrent neural network called LSTM. LSTM is a recurrent neural network that uses self-connected unbounded internal neurons called "memory cells" to store information over long time durations. Memory cells are protected by nonlinear multiplicative gates, which are employed to aid in controlling information flow through the internal states. Memory cells are organized into blocks (Figure 4), each having an input gate that allows a block to selectively ignore incoming activations; an output gate that allows a block to selectively take itself offline, shielding it from error; and a forget gate that allows cells to selectively empty their memory contents. Thus, the gates learn to protect the linear unit from irrelevant input events and error signals.

By means of gradient descent to optimize weighted connections feeding into gates as well as cells, an LSTM network can learn to control information flow. Error is back-propagated through the network in such a way that exponential decay is avoided. LSTM's learning

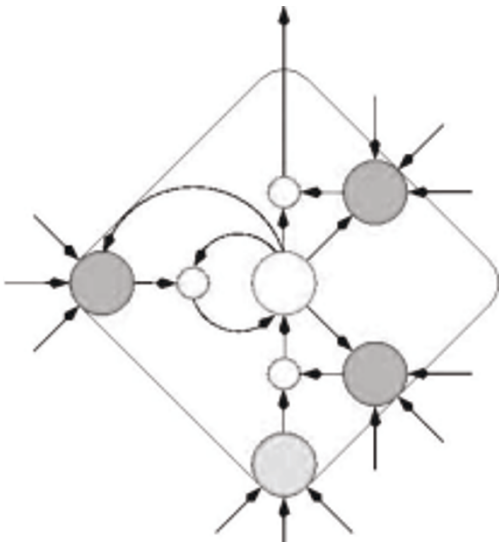

Figure 4 The memory cell of a LSTM network. The three gates (*dark-gray nodes*) collect activations from both inside and outside the block and control the cell via multiplicative units (*small white nodes*). The (*white*) node in the center of the cell is the constant error carousel (CEC), which ensures the constant error flow needed for learning long time dependencies.

algorithm is local in space and time with computational complexity per time-step and weights of O(1) for standard topologies.

4. Experiments and results

Two different emotion recognition systems have been implemented by applying the LSTM network to each representation model, namely, the MFCC model and the cochleagram model.

4.1. The dataset

As dataset for our experiments, we used the speech under simulated and actual stress (SUSAS) corpus, which has been created in the Robust Speech Processing Laboratory at Duke University under the direction of Prof. John H. L. Hansen and sponsored by the Air Force Research Laboratory. The database is partitioned into five domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 were employed to generate in excess of 16,000 utterances. SUSAS also contains several longer speech files from four Apache helicopter pilots.

For our experiments, we selected from SUSAS dataset the stress conditions' labels shown in Table 1. For each stress condition, the aircraft communication words listed in Table 2 have been considered. In Table 1. labels "Loud" and "Soft" do not properly indicate stress conditions but they are related to the voice quality. We have considered also these situations, since it is believed that the emotional content of an utterance is strongly related to voice quality. Experimental studies on listening human subjects demonstrated a strong relation between voice quality and the perceived emotion [42]. For each stress condition, 1260 utterances were considered and for each utterance, nine different speakers were considered, both male and female. For each speaker, there are two examples of the same utterance. All signals have been sampled with a sampling frequency f_s = 8 kHz.

Table 1 Stress conditions considered in our experiments.

Stress condition	Description
Angry	Simulated anger
Clear	Clearly enunciated speech
Cond70	Computer workload, high task stress
Loud	Loudly spoken speech
Soft	Soft or whispered speech

4.2. Preprocessing

Since the SUSAS dataset contains isolated sampled utterances, no segmentation was performed to extract the single pronounced words. Therefore, the preprocessing phase of our approach was used to convert each utterance into a number of voiced parts (segments). Figure 5 shows the waveform obtained from the sampled word "destination" related to an "angry" stress condition. It can be noted that the vowels are the louder parts of the waveform: in the bottom of the plot, the pronounced vowels and consonants corresponding to the signal are shown.

4.3. Representation

After the preprocessing phase, a representation for the signal was derived according to the two considered models, namely the MFCC and the Lyon cochleagram.

To obtain the first representation, we computed 24 MFCC coefficients over a number of time frames obtained by sampling each segment using a 0.015-s Hamming window at every 0.05 s. The position of the first filter of the mel filter bank was fixed at 100 mel

Table 2 Vocabulary set considered in our experiments

SUSAS vocabulary set					
SUSAS vocabulary set					
BRAKE	EIGHTY	GO	NAV	SIX	THIRTY
Change	ENTER	HELLO	NO	SOUTH	THREE
DEGREE	FIFTY	HELP	OH	STAND	WHITE
DESTINATION	FIX	HISTOGRAM	ON	STEER	WIDE
EAST	FREEZE	HOT	OUT	STRAFE	ZERO
EIGHT	GAIN	MARK	POINT	TEN	

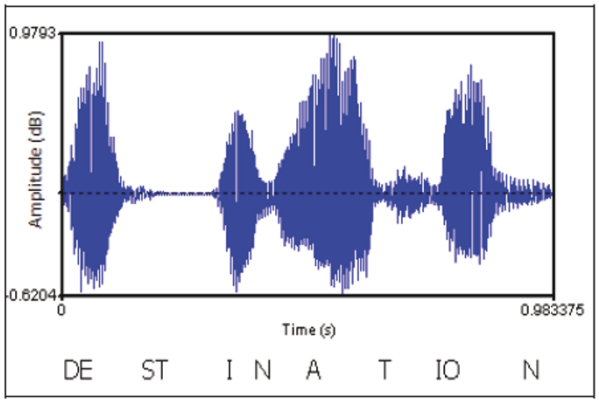

Figure 5 Sampled sound waveform of the word "destination".

and the distance between filters is of 100 mel. Figure 6 shows a 2D representation of the MFCC for the speech signal plotted in Figure 5 by means of an intensity matrix. Each row of the matrix represents the values (ranging from 0 to 4000 Hz) of a mel coefficient versus time. To compute the MFCC, we used the PRAAT software by P. Boersma and D. Weenink [43].

In the implementation of the Lyon cochlear model, we used a filter bank having 64 stages, covering a range of frequencies from 50 to 4 kHz on the same number of frames used for MFCC. Each stage is a combination of a second-order notch filter and a second-order resonance filter. Filters in the filter bank are overlapped by 25%. Figure 7 plots the Lyon cochleagram for the speech signal in Figure 5 as a matrix in which every column is a time frame and every row is a frequency band ranging from 0 to 4000 Hz. Each value (*i, j*) of this matrix provides the energy into the *i*th frequency band as a function of the *j*th time frame. To derive the Lyon cochleagram of the speech signals, we used the Auditory Toolbox developed by M. Slaney [44].

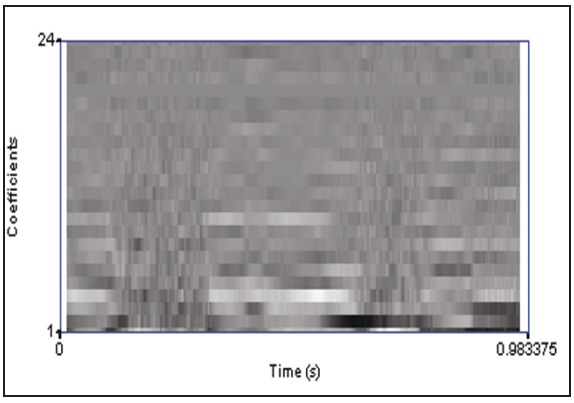

Figure 6 Mel-cepstrogram for the signal in Figure 5.

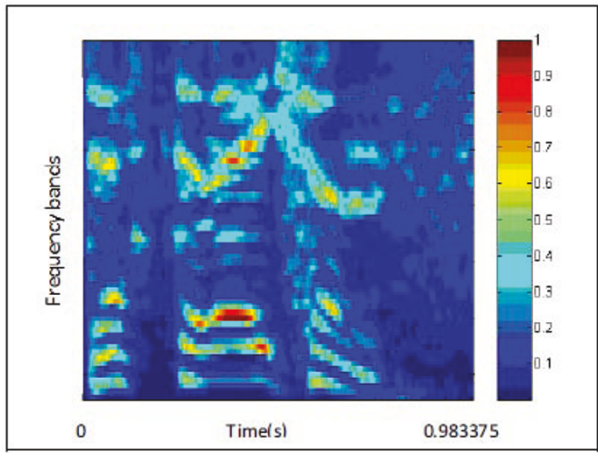

Figure 7 Lyon cochleagram for the signal of the word "destination" under angry condition.

In order to show how the Lyon cochleagram can actually model different emotions in speech, in Figure 8, we plot the Lyon cochleagram for the speech signal of the same word "destination" pronounced by the same speaker but with no emotion, i.e. neutral. Comparing Figure 7 (angry) and Figure 8 (neutral), some differences can be appreciated about the separation among different segments of the signal. Indeed, it can be seen that in Figure 7 the vowels are more separated and enhanced with high values of energy, as we expect them to be when speaking under angry condition.

4.4. Classification

Given a representation of an utterance (MFCC or cochleagram), the LSTM network was applied to perform emotion recognition, i.e. to associate the name of the emotion to each utterance. The classification process

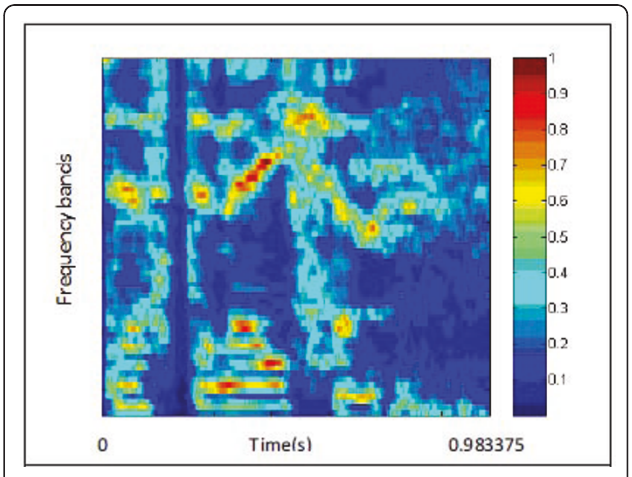

Figure 8 Lyon cochleagram for the signal of the word "destination" under neutral condition.

is defined so that the final response of the classifier is composed of a number of partial responses, as the input pattern is composed by a number of time frames extracted from the input word. Then, the final response of the classifier is an average of the responses of the LSTM network for each segment, computed as

$$r = \frac{1}{N}\sum_{i=1}^{N} r_i$$

where r_i is the response vector of the classifier to the ith segment, and N is the number of segments. The LSTM network, used as a classifier, was implemented using a different topology, depending on the signal representation adopted. A preliminary experimental session was aimed to find the best LSTM parameter configuration for each representation. Precisely, for experiments using MFCC, we found that the optimal configuration for the network topology had an input layer of size 24, a hidden layer containing 200 memory blocks (with one cell in each block), and an output layer of size 5. For experiments using the Lyon cochleagram, the optimal configuration for the network topology had an input layer of size 64, a hidden layer containing 350 memory blocks (with one cell in each block), and an output layer of size 5. In both cases, all LSTM blocks had the following activation functions:

- Logistic sigmoid in the range [-2, 2] for the input and output squashing functions of the cell;
- Logistic sigmoid in the range [0, 1] for the gates, with a gain term of 3.0.

The bias weights to the LSTM gates were initialized with positive values for the input and output gates [+0.5, +1.0, +1.5,...] and negative values for the forget gates [-0.5, -1.0, -1.5,...].

In all the experiments, online learning was used with weight update performed after every time step. Input time frames for the networks correspond to the frames (i.e. frequency content vectors) of the cochleagram computed above, for each time step. The target consists of a 5D binary vector coding one of the five possible emotion classes.

We carried out two experiments. In the first experiment, we applied the LSTM network to the MFCC. In the second experiment, we applied the LSTM network to the Lyon cochleagram. In both experiments, we considered 20 different random splits of the whole dataset into a training set (70% of the whole dataset) and test set (remaining 30% of the whole data set). Results obtained from the first experiment are shown in Table 3. The average recognition rate in this experiment was 71.5%. Results obtained from the second experiment are shown in Table 5. The average recognition rate in this experiment was 75.19%. Comparing the results obtained in the two experiments, it can be noted that the emotion classifier composed by the LSTM network applied to the Lyon cochlear representation works better for the considered dataset.

Table 3 Confusion matrix for MFCC based emotion recognition system

Response / Presented	Angry	Clear	Cond70	Loud	Soft
Angry	**73.26**	7.82	7.36	10.01	1.96
Clear	7.22	**75.41**	12.47	2.9	2
Cond70	5.67	8.29	**70.82**	10.57	4.65
Loud	4.8	8.77	8.69	**76.61**	1.13
Soft	1.66	9.23	8.24	1.02	**79.85**

A fair comparison of our approach with other existing studies was difficult, because of the different datasets and different emotion sets used in the literature. For example, in [45], the authors of the cited study report emotion recognition results from speech signals, with particular focus on extracting emotion features from the short utterances typical of interactive voice response (IVR) applications. They use a database from the Linguistic Data Consortium at University of Pennsylvania, which is recorded by eight actors expressing 15 emotions, from which they selected five emotions: hot anger, happiness, sadness, boredom, and neutral emotion. In [45], the results are reported about hot anger and neutral utterances (37 and 70%, respectively). Further by the confusion matrix, they conclude that sadness is mostly confused with boredom, happiness is mostly confused with hot anger, and neutral is mostly confused with sadness (Table 4).

A rough comparison was made by considering thestudy in [46] that used the same SUSAS dataset, but with a selection of different samples. In [46], features such as pitch, log energy, formant, mel-band energies, and MFCCs are extracted and analyzed using quadratic discriminant analysis (QDA) and SVM. With the text-independent SUSAS database, they achieved the best accuracy of 96.3% for stressed/neutral style classification and 70.1% for 4-class speaking style classification using

Table 4 Confusion matrix for Lyon cochleagram based emotion recognition system

	Response				
Presented	Angry	Clear	Cond70	Loud	Soft
Angry	**73.26**	7.82	7.36	10.01	1.96
Clear	7.22	**75.41**	12.47	2.9	2
Cond70	5.67	8.29	**70.82**	10.57	4.65
Loud	4.8	8.77	8.69	**76.61**	1.13
Soft	1.66	9.23	8.24	1.02	**79.85**

Table 5 Comparison between different emotion recognition systems

Emotion	Kwon [46]	Our best system
Clear	93.4	75.41
Angry	67.2	73.26
Loud	48.0	76.61

Gaussian SVM. The comparative results, reported in Table 5, show that our best approach (LSTM combined with Lyon cochleagram) compares favorably with the other approach.

Finally, some considerations can be done about our approach and the approaches proposed in [37,38] that apply the same LSTM network to emotion recognition from speech, but use different feature sets. Our LSTM architecture is similar to the one used in [37], but it has different parameters, due to the different feature set adopted. In general, the LSTM-RNN architecture consists of three layers: an input layer, a hidden layer, and an output layer. The number of input layer nodes corresponds to the dimension of the feature vectors. The number of hidden layer could be defined experimentally on the basis of the number of input nodes and on the performance obtained in the training phase. Therefore, we have selected as hidden layer a number ranging from 200 to 350 blocks with one cell each. For the output layer five nodes are used, corresponding to five different emotions. In [37], the size of the input layer is equal to the number of acoustic features (namely 39), and the hidden layer contains 100 memory blocks of one cell.

Our results confirm the good behavior of the LSTM network as recognizer of temporal dependencies in speech, just as in [36,38], thus encouraging its application in several tasks related to speech recognition. In particular, we observed that, despite different databases and features being used in [38], results similar to those of this study are obtained in terms of recognition accuracy, but in the case of two emotional classes. For more than two classes, results are worse than those of this study, probably because detecting valence from acoustic features is a hard task.

5. Conclusions

In this article, we have proposed a biologically inspired approach for the recognition of emotion in speech. Two different biologically plausible representations of the speech signal have been investigated, namely, the mel-scaled cepstrogram and the Lyon cochleagram. Each representation of speech has been combined with the LSTM, a biologically plausible model of artificial neural network adopted as classifier. While previous applications of the LSTM network have mainly focused on artificially generated sequence-processing tasks, this study represents one of the first efforts to apply the LSTM network in combination with biologically plausible representations of speech with the aim of emotion recognition in speech. From the experiments performed on data from the SUSAS corpus, it can be concluded that combining the LSTM classifier with the Lyon cochlear representation gives better recognition results in comparison with combining the same classifier with the traditional MFCC representation. Of course, in order to assess the validity of the presented approach, further investigations are needed on different speech datasets and for different classifier configurations. Finally, it should be noted that the presented approach can be well applied to other classification tasks involving recognition of emotional components in speech or sound signals. In particular, future study will concern the application of the proposed approach to the problem of music-style recognition.

End notes

A voiced sound involves a vibration of the vocal chords, and it is characterized by an open configuration of the vocal tract so that there is no build-up of air pressure above the glottis. This contrasts with unvoiced sound which is characterized by a constriction or closure at one or more points along the vocal tract.

Abbreviations

CEC: Constant error carousel; *SUSAS*: dual tracking computer workload; *GLR*: generalized likelihood ratio; *HMM*: hidden markov model; *IVR*: Interactive voice response; *LSTM*: Long short-term memory, *MFCC*: *mel-frequency cepstral coefficients*, *QDA*: quadratic discriminant analysis; *DLP*: speaking rate; *SUSAS*: speaking style; *SUSAS*: Speech under simulated and actual stress; *SVM*: support vector machine.

Author details

[1]Dipartimento di Informatica, Università degli Studi di Bari, Via E. Orabona 4, 70126, Bari, Italy [2]WarmPieSoft Srl, C.da S.Angelo, 72015, Fasano, Bari, Italy

Competing interests

The authors declare that they have no competing interests.

Received: 10 September 2010 Accepted: 18 July 2011
Published: 18 July 2011

References

1. J Cassel, *et al*, Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents, in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'94)*, (New York, NY, USA, 1994), pp. 413–420
2. J Cassel, *et al*, An architecture for embodied conversational character, in *Proceedings of Workshop on Embodied Conversational Characters (WECC98)*, Tahoe City, California (AAAI, ACM/SIGCHI, 1998)
3. C Breazeal, Affective interaction between humans and robots, *Advances in Artificial Life Lecture Notes in Computer Science*, ed. by Kelemen J, Sosík P (Springer, Heidelberg, 2001), pp. 582–591
4. D Ververidis, C Kotropoulos, Automatic speech recognition: resources, features and methods. Speech Commun. **48**, 1162–1181 (2006). doi:10.1016/j.specom.2006.04.003

5. FA Gers, J Schmidhuber, F Cummins, Learning to forget: continual prediction with LSTM. Neural Comput. **12**, 2451–2471 (2000). doi:10.1162/089976600300015015
6. H Hamzah, A Abdullah, A new abstraction model for biologically–inspired sound signal analyzer, in *Proceeding of 2009 IEEE Symposium on Industrial Electronics and Applications (ISIEA 2009)*, (Kuala Lumpur, Malaysia, 2009), pp. 600–605
7. L Rabiner, B-H Juang, *Fundamentals of Speech Recognition* (Prentice Hall, USA, 1993)
8. RF Lyon, CA Mead, An analog electronic cochlea. IEEE Trans Acoust Speech Sig Process. **36**, 1119–1134 (1988). doi:10.1109/29.1639
9. R Cowie, E Douglas-Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, JG Taylor, Emotion recognition in human-computer interaction. IEEE Sig Process Mag. **18**, 32–80 (2001). doi:10.1109/79.911197
10. JM Jenkins, K Oatley, NL Stein, (Eds), *Human Emotions: A Reader* (Blackwell, Malden, MA, 1998)
11. B Schuller, G Rigoll, M Lang, Hidden Markov model-based speech emotion recognition, in *Proceedings of the 2003 IEEE international conference on multimedia and expo*, (Los Alamitos, CA, USA, 2003), pp. 401–404
12. Å Abelin, J Allwood, Cross linguistic interpretation of emotional prosody, in *Proceedings of the ISCA workshop on speech and emotion* (Newcastle, Northern Ireland, 2000), pp. 110–113
13. A Tickle, English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality, in *Proceedings of the ISCA Workshop on Speech and Emotion* (Newcastle, Northern Ireland, 2000), pp. 104–109
14. KR Scherer, Adding the affective dimension: A new look in speech analysis and synthesis. in *Proceedings of International Conference on Spoken Language Processing*, 1808–1811 (1996)
15. IR Murray, JL Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature of human vocal emotion. J Acoust Soc Am. **93**, 1097–1108 (1993). doi:10.1121/1.405558
16. P Lang, The emotion probe: studies of motivation and attention. Am Psychol. **50**, 372–385 (1995)
17. H Schlosberg, Three dimensions of emotion. Psychol Rev. **61**, 81–88 (1954)
18. CE Williams, KN Stevens, Emotions and speech: some acoustical correlates. J Acoust Soc Am. **52**, 1238–1250 (1972). doi:10.1121/1.1913238
19. E Murray, JL Arnott, Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Commun. **16**, 369–390 (1995). doi:10.1016/0167-6393(95)00005-9
20. R Banse, KR Sherer, Acoustic profiles in vocal emotion expression. J Pers Soc Psychol. **70**, 614–636 (1996)
21. Y Sagisaka, N Campbell, N Higuchi, (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (Springer, New York, NY, 1997)
22. RB Monsen, AM Engebretson, R Vemula, Indirect assessment of the contribution of subglottal air pressure and vocal fold tension to changes in fundamental frequency in English. J Acoust Soc Am. **64**, 65–80 (1978). doi:10.1121/1.381957
23. IR Titze, *Principles of Voice Production* (Prentice-Hall, London, 1993)
24. DB Fry, Duration and intensity as physical correlates of linguistic stress. J Acoust Soc Am. **30**, 765–769 (1955)
25. P Lieberman, Some acoustic correlates of word stress in American-English. J Acous Soc Am. **32**, 451–454 (1960). doi:10.1121/1.1908095
26. K Maekawa, Phonetic and phonological characteristics of paralinguistic information in spoken Japanese, in *Proceedings of the International Conference on Spoken Language Processing* (1998)
27. AMC Sluijter, VJ van Heuven, Spectral balance as an acoustic correlate of linguistic stress. J Acoust Soc Am. **100**, 2471–2485 (1996). doi:10.1121/1.417955
28. T New, S Foo, L De Silva, Speech emotion recognition using hidden Markov models. Speech Commun. **41**, 603–623 (2003). doi:10.1016/S0167-6393(03)00099-2
29. L Fu, X Mao, L Chen, Speaker independent emotion recognition based on svm/hmms fusion system, in *Proceedings of International Conference on Audio, Language and Image Processing (ICALIP 2008)*, 61–65 (2008)
30. F Dellaert, T Polzin, A Waibel, Recognizing emotion in speech, in *Proceedings of International Conference on Spoken Language Processing*, 1970–1973 (1996)
31. VA Petrushin, Emotion in speech: recognition and application to call centers, in *Proceedings of Artificial Neural Networks in Engineering*, 7–10 (1999)
32. LS Chen, Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction, PhD Paper, Department of Electrical Engin, University of Illinois at Urbana-Champaign, 2000
33. CA Buscicchio, P Gorecki, L Caponetti, Emotion recognition in speech signals, in *Proceedings of 16th International Symposium on Foundations of Intelligent Systems (ISMIS 2006)* (Bari, Italy, 2006), pp. 38–46
34. HJM Steeneken, JHL Hansen, Speech under stress conditions: overview of the effect of speech production and on system performance, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-99)* (Phoenix, Arizona, March, 1999), pp. 2079–2082
35. S Hochreiter, J Schmidhuber, Long short-term memory, in *Neural Comput.* **9**, 1735–1780 (1997). doi:10.1162/neco.1997.9.8.1735
36. DE Graves, N Beringer, J Schmidhuber, in *Biologically plausible speech recognition with LSTM neural nets*, ed. by Ijspeert AJ, Murata M, Wakamiya N BioAdit 2004. Lecture Notes in Computer Science, vol. 3141 (Springer, New York, 2004), pp. 127–136.
37. M Wöllmer, F Eyben, S Reiter, B Schuller, C Cox, E Douglas-Cowie, R Cowie, Abandoning emotion classes–towards continuous emotion recognition with modelling of long-range dependencies, in *Proceedings of 9th Interspeech Conference* (Brisbane, Australia, 2008), pp. 597–600
38. M Wöllmer, F Eyben, B Schuller, E Douglas-Cowie, R Cowie, Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM net-works, in *Proceedings of Interspeech* (Brighton, UK, 2009), pp. 1595–1598
39. M Wöllmer, F Eyben, J Keshet, A Graves, B Schuller, G Rigoll, Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks, in *Proceedings of ICASSP* (Taipei, Taiwan, 2009)
40. A Von Brandt, Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test, in *Proceedings of ICASSP* (Boston, MA, 1983), pp. 1017–1020
41. MD Mauk, DV Buonomano, The neural basis of temporal processing. Annu Rev Neurosci. **27**, 307–340 (2004). doi:10.1146/annurev.neuro.27.070203.144247
42. C Gobl, AN Chasaide, The role of voice quality in communicating emotion, mood and attitude. Speech Commun. **40**, 189–212 (2003). doi:10.1016/S0167-6393(02)00082-1
43. Praat homepage: http://www.fon.hum.uva.nl/praat
44. M Slaney, *Auditory Toolbox, Version 2. Technical Report 1998-010*, (Interval Research Corporation, 1998) http://www.slaney.org/malcolm/pubs.html
45. S Yacoub, S Simske, X Lin, J Burns, Recognition of emotions in interactive voice response systems, in *Proceedings of Eurospeech 2003* (Geneva, Switzerland, 2003), pp. 729–732
46. O-W Kwon, K-L Chan, J Hao, T-W Lee, Emotion recognition by speech signals, in *Proceedings of Eurospeech 2003* (Switzerland, 2003), pp. 125–128

doi:10.1186/1687-6180-2011-24
Cite this article as: Caponetti *et al.*: **Biologically inspired emotion recognition from speech.** *EURASIP Journal on Advances in Signal Processing* 2011 **2011**:24.