Real-Time Emotion Recognition from Speech Using Echo State Networks

Stefan Scherer, Mohamed Oubbati, Friedhelm Schwenker, and Günther Palm

Institute of Neural Information Processing, Ulm University, Germany
{stefan.scherer,mohamed.oubbati,friedhelm.schwenker,
guenther.palm}@uni-ulm.de

Abstract. The goal of this work is to investigate real-time emotion recognition in noisy environments. Our approach is to solve this problem using novel recurrent neural networks called echo state networks (ESN). ESNs utilizing the sequential characteristics of biologically motivated modulation spectrum features are easy to train and robust towards noisy real world conditions. The standard Berlin Database of Emotional Speech is used to evaluate the performance of the proposed approach. The experiments reveal promising results overcoming known difficulties and drawbacks of common approaches.

1 Introduction

The present innovations in affective computing aim to provide simpler and more natural interfaces for human-computer interaction applications. Detecting and recognizing the emotional status of a user is important in designing and developing efficient and productive human-computer interaction interfaces [2]. The efficiency gain is well founded on the fact that in healthy human to human interaction emotion is essential in every bit of communication. For example, while explaining something to another person, one could communicate understanding with a simple smile, with no need to say "I understand, what you are telling me" [15]. Hence, emotion analysis and processing is a multi-disciplinary topic, which has been emerging as a rich research area in recent times [2,4,6,13,18]. The visual cues, such as facial expressions and hand gestures are the natural indicators of emotions. However, these require additional hardware and computational resources for processing. Alternatively, speech can be used for emotion recognition which is not only simple to process, but can also be incorporated into the existing speech processing applications [2,6,17]. Most commonly used features are pitch, energy and speech spectral based features [13]. In this work, a novel approach based on long term modulation spectrum of speech is used to detect the emotions close to real-time using a recurrent neural network called echo state network (ESN).

One of the main issues in designing an automatic emotion recognition system is the selection of the features that can represent the corresponding emotions. In [12], pitch and linear predictive coding (LPC) features were used as input

L. Prevost, S. Marinai, and F. Schwenker (Eds.): ANNPR 2008, LNAI 5064, pp. 205–216, 2008.

to an artificial neural network (ANN). After detecting the start and end points of the utterances, a 300 dimensional vector was used, which resulted in classification rates of around 50% detecting eight different emotions. In earlier work multi classifier systems (MCS) were trained with three feature types, comprising modulation spectrum, as in this work, relative spectral transform - perceptual linear prediction (RASTA-PLP), and perceived loudness features, in a MCS to recognize seven different emotions with an accuracy of more than 70% [18]. The Mel Frequency Cepstral Coefficients (MFCC) based features were used in [11], which were obtained with a window of 25 ms sampled every 10 ms. The Hidden Markov Model (HMM) was then used for training each of the four targeted emotions. After training a combination of all the phoneme class based HMMs on the TIMIT database, for each of the emotions, the classification performance reached around 76%. In [3], k-nearest neighbor (KNN) algorithm was applied to classify four emotions, which resulted in 65 % accuracy. The pitch based statistics, such as contour of pitch, maximum, minimum, and slope were considered as features. Broadly speaking, differences in the various approaches arise from the overall goal (recognizing single vs. multiple emotions), the specific features used, and the classification framework. The anger vs. neutral emotion classification was studied, particularly in the context of interactive voice response systems with specific application to call centers in [19]. 37 prosody features related to pitch, energy, and duration were used as features, and for classification neural networks, support vector machines (SVM), and KNN were applied. With the most significant feature set of 19 features, the best recognition accuracy about 90% was achieved using SVMs. In another study, agitation vs. calm emotion classification was performed with KNN, ANN, and set of experts [14]. “Agitation” included happiness, anger, and fear, and “calm” comprised neutral, and sadness emotional states. Pitch, energy, speaking rate, formants, and their bandwidths were used as features, which resulted in an accuracy of 77%.

However, commonly used features and classifiers are sensitive towards noise. In this work, a system overcoming these issues is targeted. Furthermore, classifiers still require time to classify the utterances as they rely on statistics of the features and are computationally intensive. Here we use a special characteristic of long term modulation spectrum, which reflects syllabic and phonetic temporal structures of speech [5,7]. Recently, a novel recurrent neural network (RNN) called echo state network (ESN) is developed. An ESN has an easy training algorithm, where only the output weights are to be adjusted. The basic idea of ESN is to use a Dynamic Reservoir (DR), which contains a large number of sparsely interconnected neurons with non-trainable weights. The previously mentioned features are used as inputs to an ESN classifier. Since, the only weights that need to be adjusted in an ESN are the output weights, the training is not computationally expensive using the direct pseudo inverse calculation instead of gradient descent training. The performance is close to real-time, as the decisions are made on short-segments of the signal (100 ms), rather than over the entire utterance. This is of great advantage since emotions are constantly changing and aggregating statistics of pitch or other similar features may not suffice [17]. An

example of a scenario where emotions change rapidly can be found in [15]. In this scenario a tennis player feels a piercing pain in his lower back and he first turns around clenching his fist and feeling angry, but as he sees that a woman in a wheelchair hit him his feelings changed to sadness and sympathy. This small example illustrates the possible rapid changes of how we value situations. Therefore, it is necessary to build a system that does not need to aggregate information over several seconds, but is able to classify emotions close to real time. In this work we present a feature extraction system that extracts after a lead time of 400 ms feature vectors with a frequency of 25 Hz, which is sufficient for emotion recognition in many applications.

The paper is organized and presented in four sections: Section 2 gives an overview of the database used for experiments, Sect. 3 describes the feature extraction, Sect. 4 introduces the echo state networks used for classifictation, Sect. 5 presents the experiments and results, and finally Sect. 6 concludes.

2 Database Description

The Berlin Database of Emotional Speech is used as a test bed for our approach. This corpus is a collection of around 800 utterances spoken in seven different emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral [1]. The database is publicly available at `http://pascal.kgw.tu-berlin.de/emodb/`. Ten professional actors (five male and five female) read the predefined utterances in an anechoic chamber, under supervised conditions. The text was taken from everyday life situations, and did not include any emotional bias. The utterances are available at a sampling rate of 16 kHz with a 16 bit resolution and mono channel. A human perception test to recognize various emotions with 20 participants resulted in a mean accuracy of around 84% [1].

3 Feature Extraction

Short term analysis of the speech signal, such as extracting spectral features from frames not more than several milliseconds, dominates speech processing for many years. However, these features are strongly influenced by environmental noise and are therefore unstable. In [8], it is suggested to use the so called modulation spectrum of speech to obtain information about the temporal dynamics of the speech signal to extract reliable cues for the linguistic context. Since emotion in speech is often communicated by varying temporal dynamics in the signal the same features are used to classify emotional speech in the following experiments [17].

The proposed features are based on long term modulation spectrum. In this work, the features based on slow temporal evolution of the speech are used to represent the emotional status of the speaker. These slow temporal modulations of speech emulate the perception ability of the human auditory system. Earlier studies reported that the modulation frequency components from the range

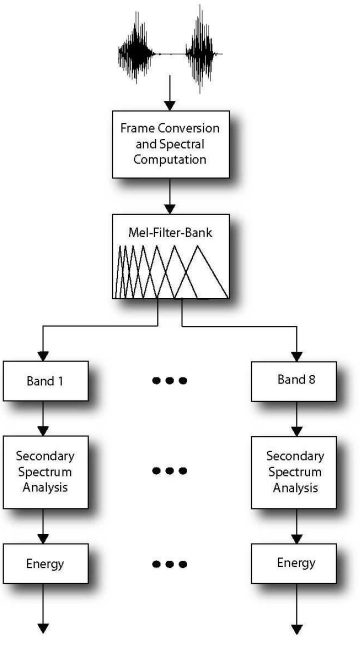

Fig. 1. Schematic description for feature extraction

between 2 and 16 Hz, with dominant component at around 4 Hz , contain important linguistic information [5,7,10]. Dominant components represent strong rate of change of the vocal tract shape. This particular property, along with the other features has been used to discriminate speech and music [16]. In this work, the proposed features are based on this specific characteristic of speech, to recognize the emotional state of the speaker.

The block diagram for the feature extraction for a system to recognize emotions is shown in Fig. 1. The fast Fourier transform (FFT) for the input signal $x(t)$ is computed over N points with a shift of n samples, which results in a $\frac{N}{2}$ dimensional FFT vector. Then, the Mel-scale transformation, motivated by the human auditory system, is applied to these vectors. The Mel-filter bank with eight triangular filters $H_i[k]$, is defined by:

$$H_i[k] = \begin{cases} \frac{2(k-b_i)}{(d_i-b_i)(c_i-b_i)} & b_i \leq k \leq c_i \\ \\ \frac{2(d_i-k)}{(d_i-b_i)(d_i-c_i)} & c_i \leq k \leq d_i \end{cases}, \tag{1}$$

where $i = 1, ..., 8$ indicates the index of the i-th filter. b_i and d_i indicate the frequency range of filter H_i and the center frequency c_i is defined as $c_i = (b_i + d_i)/2$. These ranges are equally distributed in the Mel-scale, and the corresponding frequencies b_i and d_i are listed in Table 1. For $k < b_i$ and $k > d_i$ $H_i[k] = 0$.

Table 1. Start and end frequencies of the triangular Mel-filters

Band	Start Freq. (Hz)	End Freq. (Hz)
1	32	578
2	257	964
3	578	1501
4	966	2217
5	1501	3180
6	2217	4433
7	3180	6972
8	4433	8256

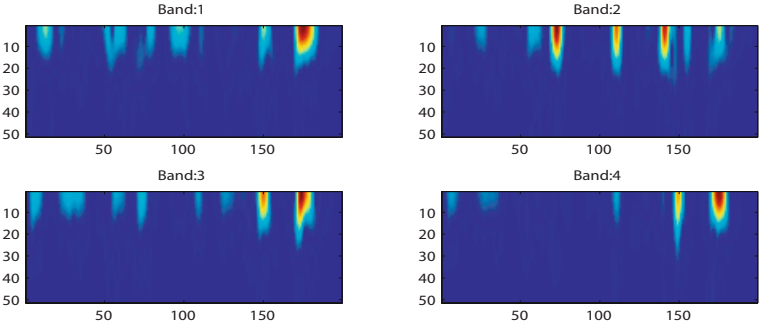

Fig. 2. Modulation spectrum for the first four bands of a single angry utterance. The x-axis represents the time scale, in frames and the y-axis, the frequency in Hz.

For each of the bands, the modulations of the signal are computed by taking FFT over the P points, shifted by p samples, resulting in a sequence of $\frac{P}{2}$ dimensional modulation vectors. Most of the prominent energies can be observed within the frequencies between 2 - 16 Hz. Figure 2 illustrates the modulation spectrum based energies for a single angry utterance, for the values $N = 512$, $n = 160$, $P = 100$ and $p = 1$ for the first four bands. For the classification task following values were used: $N = 1600$, $n = 640$, $P = 10$, $p = 1$. Since the signal is sampled with 16 kHz, N corresponds to 100 ms and n to 40 ms resulting in a feature extraction frequency of 25 Hz. According to the window size P a lead time of 400 ms is necessary. Therefore, one feature vector in the modulation spectrum takes 400 ms into account with an overlap of 360 ms, due to p.

4 Echo State Networks

Feed forward neural networks have been successfully used to solve problems that require the computation of a static function, i.e. a function whose output depends only upon the current input. In the real world however, many problems cannot be solved by learning a static function because the function being computed may produce different outputs for the same input if it is in different states. Since

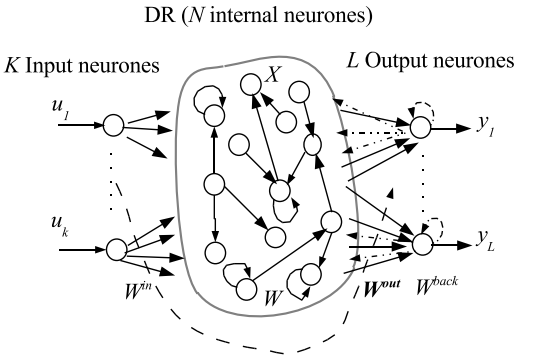

Fig. 3. Basic architecture of ESN. Dotted arrows indicate connections that are possible but not required.

expressing emotions is a constantly changing signal, emotion recognition falls into this category of problems. Thus, to solve such problems, the network must have some notion of how the past inputs affect the processing of the present input. In other words, the network must have a memory of the past input and a way to use that memory to process the current input. This limitation can be rectified by the introduction of feedback connections in the network. The class of Neural Networks which contain feedback connections are called RNNs. In principle RNNs can implement almost arbitrary sequential behavior, which makes them promising for adaptive dynamical systems. However, they are often regarded as difficult to train. Using ESNs only two steps are necessary for training: First, one forms a DR, with input neurons and input connections, which has the echo state property. The echo state property says: "if the network has been run for a very long time, the current network state is uniquely determined by the history of the input and the (teacher-forced) output." [9]. According to experience, it is better to ensure that the internal weight matrix has maximum eingenvalue $|\lambda_{max}| < 1$. Second, one attaches output neurons to the network and trains suitable output weights.

As presented in (Fig. 3), we consider a network with K inputs, N internal neurons and L output neurons. Activations of input neurons at time step n are $U(n) = (u_1(n), \ldots, u_k(n))$, of internal units are $X(n) = (x_1(n), \ldots, x_N(n))$, and of output neurons are $Y(n) = (y_1(n), \ldots, y_L(n))$. Weights for the input connection in a (NxK) matrix are $W^{in} = (w_{ij}^{in})$, for the internal connection in a (NxN) matrix are $W = (w_{ij})$, and for the connection to the output neurons in an L x $(K + N + L)$ matrix are $W^{out} = (w_{ij}^{out})$, and in a (NxL) matrix $W^{back} = (w_{ij}^{back})$ for the connection from the output to the internal units.

The activation of internal and output units is updated according to:

$$X(n+1) = f(W^{in}U(n+1) + WX(n) + W^{back}Y(n)) \tag{2}$$

where $f = (f_1, \ldots, f_N)$ are the internal neurons output sigmoid functions. The outputs are computed according to:

$$Y(n+1) = f^{out}(W^{out}(U(n+1), X(n+1), Y(n))) \tag{3}$$

where $f^{out} = (f_1^{out}, \ldots, f_L^{out})$ are the output neurons output sigmoid functions. The term $(U(n+1), X(n+1), Y(n))$ is the concatenation of the input, internal, and previous output activation vectors. The idea of this network is that only the weights for connections from the internal neurons to the output (W^{out}) are to be adjusted.

Here we present briefly an off-line algorithm for the learning procedure:

1. Given I/O training sequence $(U(n), D(n))$
2. Generate randomly the matrices (W^{in}, W, W^{back}), scaling the weight matrix W such that it's maximum eingenvalue $|\lambda_{max}| \leq 1$.
3. Drive the network using the training I/O training data, by computing

$$X(n+1) = f(W^{in}U(n+1) + WX(n) + W^{back}D(n)) \tag{4}$$

4. Collect at each time the state $X(n)$ as a new row into a state collecting matrix M, and collect similarly at each time the sigmoid-inverted teacher output $tanh^{-1}D(n)$ into a teacher collection matrix T.
5. Compute the pseudo inverse of M and put

$$W^{out} = (M^{+}T)^{t} \tag{5}$$

t: indicates transpose operation.

For exploitation, the trained network can be driven by new input sequences and using the equations (2) and (3).

5 Experiments and Results

All the experiments were carried out on the German Berlin Database of Emotional Speech, which is described in Sect. 2. The utterances comprising anger and neutral are specifically used with regard to the most important task in call center applications, where it is necessary to recognize angry customers for the system to react properly.

In the first experiment especially the real-time recognition capability of the ESN is tested. After thorough tuning the following parameters revealed optimal results. A randomly initialized network of 500 neurons was used. The connectivity within the network was 0.5 which indicates that 50% of the connections were set within the network. Additionally, the spectral width λ_{max} was set to 0.2. Figure 4 (a) shows the cumulative sums over frame wise decisions of three different utterances taken from the first fold of the 10 fold cross validation experiment. The vertical bars represent borders of single utterances taken from the database. The ESN needs some time to adapt to the new portrayed emotion

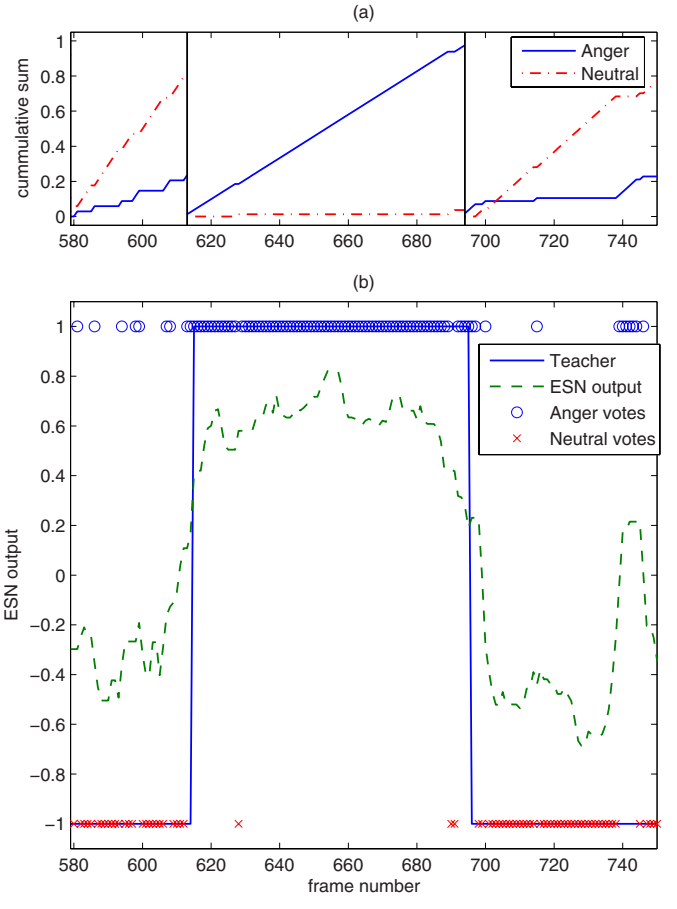

Fig. 4. Example results of ESN generalization behavior: (a) cumulative sum over frame wise decisions; (b) median filtered ESN output with superimposed plotted frame wise decisions and teacher signal

in the following utterance. However, after only a few frames the ESN achieves adoption and recognizes the correct emotion fast. The output of the ESN counts as a vote for anger if the sign of the output is + and a vote for neutral if it is −. It is seen that the cumulative sum of the correct emotion is above the other only a few frames after an emotional shift and hardly any errors are made in most of the cases. At every point in time a possible shift from neutral to angry and vice versa is possible. Only once in the 10 folds of the cross validation the correct emotion is overruled by around 52% of the votes, leading to an accuracy of more than 99%. In every other case the correct emotion wins at least at the end of an utterance, most of the cases behave similarly to the three in Fig. 4 (a). In Fig. 4 (b) the dashed line corresponds to the median filtered ESN output signal. The median filter used had the magnitude 10. However, the circles and

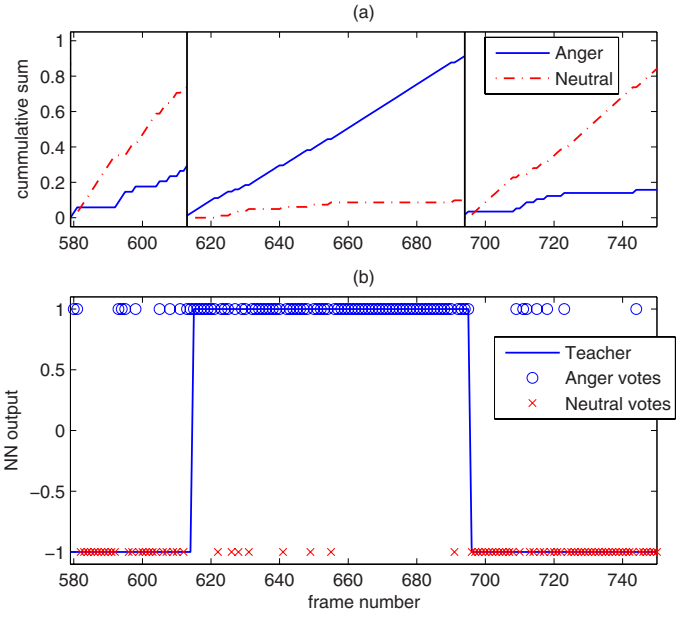

Fig. 5. Example results of NN generalization behavior: (a) cumulative sum over frame wise decisions; (b) frame wise decisions with superimposed plotted teacher signal

crosses correspond to the frame wise decisions not the median filtered output in order to show the real-time performance of the network. Additionally, the plot of the ESN output is superimposed by the teacher signal. Each circle or cross that does not lie on the solid line resembles an error of the network on a single frame. These errors may occur, but oscillating outputs could just be filtered out and only if a series of the same decisions follow each other an emotion will be recognized in future applications.

In Fig. 5 (a) the cumulative sums over the frame wise decisions of a simple NN classifier are shown. In most of the cases the NN errs more often than the ESN. This may be the result of the ESN's capability to take earlier frames into account. Due to the recurrent architecture of the ESN it is possible to "keep several frames in mind". In a similar way HMMs are capable to process sequences of frames. The cumulative sums of the wrong emotion overruled the correct emotion in the 10 fold cross validation 6 times using the NN classifier. Furthermore, the calculation of the nearest neighbor is far more computationally expensive as the output for a single frame of the ESN. In Fig. 5 (b) again circles or crosses not lying on the line resembling the teacher signal count as errors. To be able to calculate the outputs in real-time it would be necessary to reduce the search space for the NN classifier. For example, by using a clustering method such as Learning Vector Quantization (LVQ) or K-Means. However, using these methods may result in more errors.

Table 2. Frame wise error rate of the two classifiers according to differing conditions of noise

Type of Noise	Frame wise error rates	
	ESN	NN
no noise	0.15	0.21
coffee machine	0.14	0.23
office	0.16	0.22
vacuum cleaner	0.16	0.23
inside car	0.18	0.26
city/street	0.18	0.27

In a second experiment we added different amounts of noise to the audio signals in order to check whether the ESN or the NN are capable of dealing with noise and how the classification performance develops. Theoretically the modulation spectrum features should be quite stable towards noise, since only voice relevant frequencies pass through the Mel filtering. Additionally, it is possible to compensate noise using ESNs [9]. Table 2 shows the recognition results adding noise to the audio signal. The error rates correspond to the average rate of misclassified frames. In the first row no additional noise was added. In the following rows the amount of noise slowly increases from a quiet coffee machine up to noise recorded in Helsinki on a sidewalk. The noise of the coffee machine is mostly due to dripping water. The office environment corresponds to people typing, chatting in the background, and copying some papers. The vacuum cleaner is a very constant but loud noise. Inside the car mostly the engine of the own car is heard. The last type of noise comprises passing trucks, cars, motorbikes, and people passing by. All the noise was added to the original signal before feature extraction. The audio material was taken from the homepage of the "Freesound Project"[1]. It is seen that the results using modulation spectrum based features are quite robust using both classifiers. However, the recognition stays more stable using the ESN, which confirms the abilities of the ESN to compensate noise to a certain amount.

6 Conclusions

This paper presented an approach towards recognizing emotions from speech close to real-time. Features motivated by the human auditory system were used as input for an ESN. The real-time recognition performance of the network in one of the most important tasks in automatic call centers is impressive. Furthermore, the utilized features as well as the ESN are very stable towards noise of different types, like cars, office environment, or a running vacuum cleaner. The data was tested using a standard emotional speech database. However, the recognition task was rather limited and could be extended further to recognize

[1] Data is freely available: `http://freesound.iua.upf.edu/`

various other emotions such as happiness, fear, disgust, and sadness, and towards the classification of real-world data. In earlier work, these emotions were recognized using additional feature types and multi classifier systems [18]. In this work emotions can not be recognized in real-time, which is a major drawback as it was illustrated in an example in this paper. However, this could be improved using ESNs. These issues will be studied in the future.

Acknowledgements

This work is supported by the competence center Perception and Interactive Technologies (PIT) in the scope of the Landesforschungsschwerpunkt project: "Der Computer als Dialogpartner: Perception and Interaction in Multi-User Environments" funded by the Ministry of Science, Research and the Arts of Baden-Württemberg.

References

1. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Proceedings of Interspeech 2005 (2005)
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)
3. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. In: Proceedings of ICSLP, pp. 1970–1973 (1996)
4. Devillers, L., Vidrascu, L., Lamel, L.: Challanges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422 (2005)
5. Drullman, R., Festen, J., Plomp, R.: Effect of reducing slow temporal modulations on speech reception. Journal of the Acousic Society 95, 2670–2680 (1994)
6. Fragopanagos, N., Taylor, J.G.: Emotion recognition in human-computer interaction. Neural Networks 18, 389–405 (2005)
7. Hermansky, H.: Auditory modeling in automatic recognition of speech. In: Proceedings of Keele Workshop (1996)
8. Hermansky, H.: The modulation spectrum in automatic recognition of speech. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (1997)
9. Jaeger, H.: Tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the echo state network approach. Technical Report 159, Fraunhofer-Gesellschaft, St. Augustin Germany (2002)
10. Kanederaa, N., Araib, T., Hermansky, H., Pavele, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communications 28, 43–55 (1999)
11. Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S.S.: Emotion recognition based on phoneme classes. In: Proceedings of ICSLP 2004 (2004)
12. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion recognition in speech using neural networks. Neural Computing and Applications 9, 290–296 (2000)

13. Oudeyer, P.-Y.: The production and recognition of emotions in speech: features and algorithms. International Journal of Human Computer Interaction 59(1-2), 157–183 (2003)
14. Petrushin, V.: Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering (1999)
15. Picard, R.W.: Affective Computing. MIT Press, Cambridge (2000)
16. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Proceedings of ICASSP, vol. 1, pp. 1331–1334 (1997)
17. Scherer, K.R., Johnstone, T., Klasmeyer, G.: Affective Science. In: Handbook of Affective Sciences - Vocal expression of emotion, pp. 433–456. Oxford University Press, Oxford (2003)
18. Scherer, S., Schwenker, F., Palm, G.: Classifier fusion for emotion recognition from speech. In: Proceedings of Intelligent Environments 2007 (2007)
19. Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of emotions in interactive voice response systems. In: Proceedings of Eurospeech 2003 (2003)