Real-Time Emotional Speech Processing for Neurorobotics Applications

C. M. Thibeault[1,2,3,*], O. Sessions[3], P. H. Goodman[3,4], and F. C. Harris Jr.[2,3]

[1] *Department of Electrical and Biomedical Engineering, University of Nevada, Reno. Reno, NV.*
[2] *Department of Computer Science, University of Nevada, Reno. Reno, NV.*
[3] *Brain Computation Lab, University of Nevada, Reno. Reno, NV.*
[4] *Department of Internal Medicine and Program in Biomedical Engineering, University of Nevada, Reno. Reno, NV.*

Abstract

The ability for humans to understand and process the emotional content of speech is unsurpassed by simulated intelligent agents. Beyond the linguistic content of speech are the underlying prosodic features naturally understood by humans. The goal of emotional speech processing systems is to extract and classify human speech for these so called paralinguistic elements. Presented here is a proof-of-concept system designed to analyze speech in real-time for coupled interactions with spiking neural models. Based on proven feature extraction algorithms, the resulting system provides two interface options to running simulations on the NeoCortical Simulator. Some basic tests using new recordings as well as a subset from a published emotional database were completed with promising results.

1 Introduction

Much of human communication is not in what is said but how it is spoken. These subtle changes in emotion that exist beyond the linguistic aspects and the perceptual ability to interpret them is fundamental to speech. There have been many studies aimed at parameterizing and classifying such emotions. While many of these investigations have taken advantage of advances in neural networks as well as statistical and probabilistic classification mechanisms, the authors are unaware of such studies employing biologically realistic neural networks. These so called spiking networks strive to model neurons and neural assemblies with as much biological realism as is computationally feasible. The combination of biological neural networks and high-level speech processing proposes a unique opportunity to explore some of the possible neural mechanisms behind emotional expression. Additionally, these networks may aid in the creation of more successful emotional classification tools. This project is a first step towards the combination of emotional speech processing (ESP) and computational neuroscience in a real-time architecture that can be easily modified and extended.

*Corresponding Author. Corey@cmthibeault.com

1.1 Emotional Speech Processing

There has been a wealth of research on the extraction of emotional information from speech. Unfortunately, this work has yet to identify a standard set of features that completely identify a signal. The most common features found in the literature deal with pitch. Additionally, the energy, rate, and frequency components of a signal have been employed with varying rates of success.

Acoustic Properties

Of the acoustic properties researched by the community, pitch appears to be one of the more popular. A product of the tension of the vocal-cords and the corresponding air pressure, pitch changes over time show a direct relation to some basic emotions (e.g. anger and happiness) [1]. Ververidis *et al.* [2], presented a collection of how many acoustic features, including pitch, correspond to some of the basic emotions. Similarly, the intensity of a signal can be used to classify emotional content.

Although there is considerable empirical evidence supporting the classification of a speech signal based on the acoustic features, there is also wide variation between these studies. Additionally, it has been argued that these features really only provide information about the arousal state of the speaker, rather than their true emotional state [1].

1.2 Previous Work

Previous studies on emotional speech recognition use methods ranging from frequency analysis, segmental analysis or prosodic features, as well as analysis of the signal intensity. For comprehensive reviews of the current literature of these methods see [2] and [3]. From these reviews, it becomes obvious that the concept of extracting emotional information independent

of the linguistic content is not new. Additionally, there have been several applications identified for these types of classification systems. Some examples are presented in the discussion of Section 5 below.

1.3 Neurorobotics and The NeoCortical Simulator

The NeoCortical Simulator (NCS) is a spiking neural simulator developed at the University of Nevada, Reno. NCS models large-scale networks of conductance-based integrate and fire neurons. Developed with an emphasis on both biological realism and high-performance, NCS provides researchers a unique environment for computational neuroscience. For a review of NCS and the other major computational neuroscience simulators see [4].

A feature unique to NCS is its network interface. There are a number of tools developed and in development for interacting with a running NCS simulation. These interfaces have been exploited for a number of unique interactions. One of particular interest is neurorobotics [5, 6]. Neurorobotics aims at developing combinations of biologically realistic neural simulations with robotic agents and human participants in closed-loop configurations.

1.4 Contribution of This Paper

This paper presents an emotional speech processing system offering both real-time performance and a simple programming interface, as well as a direct interface to a biologically realistic neural simulator. The remainder of the paper continues with Section 2, describing the overall system design. Section 3 describes the initial testing completed over the course of development. With Section 4 giving the results of those tests. Finally, Section 5 concludes with a brief discussion of its applications as well as future work.

2 ESP System

The overarching goals of this project were to develop a complete emotional speech processing system that could not only perform in soft real-time but could be modified and extended by users with limited programming skills. The MATLAB environment was chosen mostly for its ease of use. Its interpreted language processing can often be a disadvantage, however for this project some basic optimizations facilitated real-time performance.

The system consisted of three processing sections: Audio Capture, Signal Processing and Data Communication. The speech processing system is diagrammed in Figure 1 and explained further below.

2.1 Data Capture

The audio stream is captured using the Data Acquisition Toolbox for MATLAB, developed by The Mathworks Inc. Data is captured in 1 second blocks before being sent to a custom callback function. The Analysis begins with an extraction of the significant segments in the speech signal. There are many different methods for segmentation, see [7, 8], here, a simple speech energy calculation is used.

The speech energy of a signal is an average sum of squares calculation that can represent the overall energy in a window of speech. The speech energy can be employed for distinguishing actual speech signals from background noise. Although not employed in this system, the speech energy can additionally be used to help classify the emotional arousal level of the speaker. It can be calculated by Equation (1).

$$E(m) = \frac{1}{N} \sum_{n=0}^{N} |x(n)|^2 \tag{1}$$

This calculation is completed on 20ms windows of data. The results are stored and compared to a user defined threshold. When that threshold is reached the system will begin extracting sampled data until the threshold is crossed again. Finally, the extracted window of data is sent to the signal processing blocks. Any left-over data will be retained and attached to subsequent recordings.

2.2 Signal Processing

The feature extraction begins with a calculation of the mean and range of the raw intensity values. The segment is then run through an autocorrelation algorithm and the fundamental frequencies of 20ms windows is computed. The window is stepped by 10ms allowing overlap of pitch calculations.

Autocorrelation Analysis

The autocorrelation calculation has been shown to faithfully extract the fundamental frequency (pitch) of speech signals. In general the autocorrelation calculation is not appropriate for any continuous function. However, by taking piece-wise windows of the continuous signal, the stationarity assumption can be applied to the individual window [9]. This assumption allows the use of sample autocorrelation calculations on continuous signals and illustrates its appropriateness for this application.

The autocorrelation of a windowed signal, $x(n)$ with N samples, can be defined as Equation (2) [9, 10].

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n)x(n+m) \tag{2}$$

Requiring programmatic loops, this can be a computationally expensive calculation. That cost can be reduced by considering the calculation as an ordinary convolution. The autocorrelation can then be

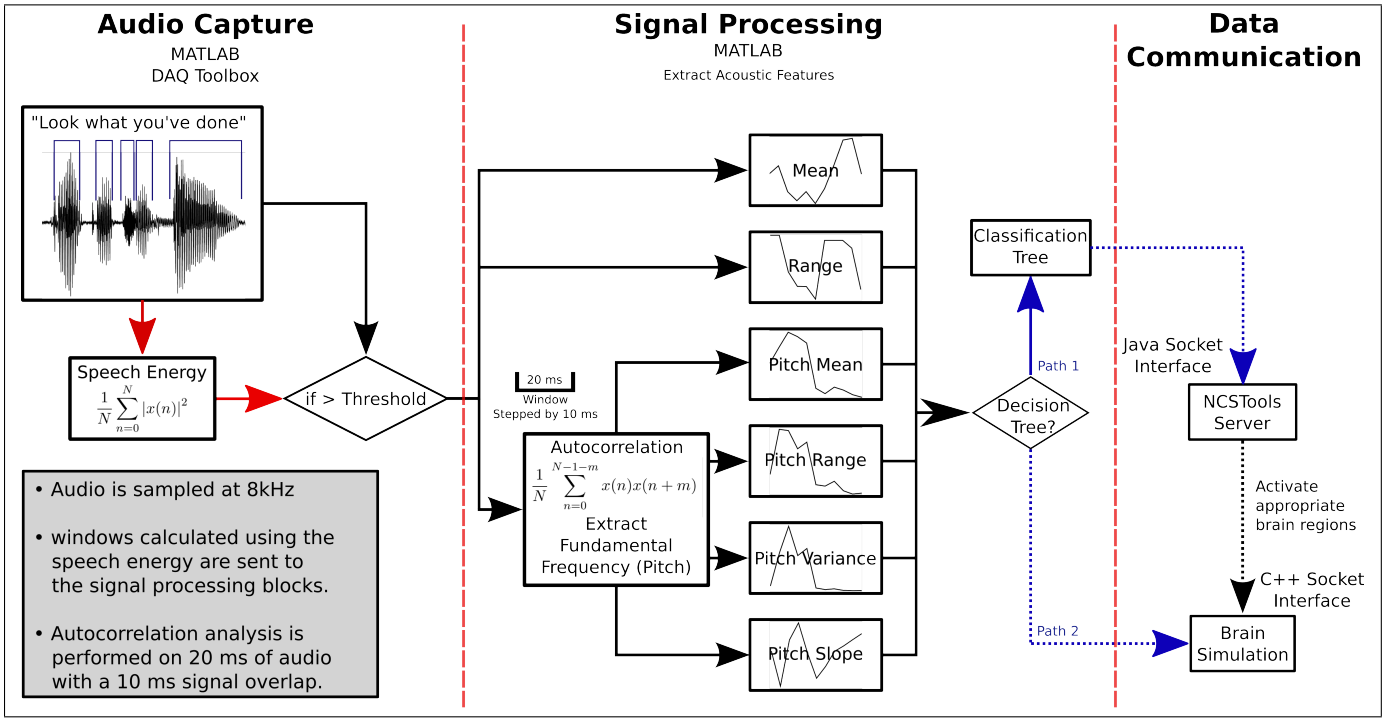

Figure 1: Speech Processing System

computed using the periodogram spectrum defined as Equation (3) [9].

$$S(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n)e^{-j\omega n}\right|^2 \tag{3}$$

This is sometimes referred to as the short-time spectrum. As defined by the Wiener-Khinchin theorem the combination of $S(\omega)$ and $R(m)$ are a simple Fourier-Transform pair. Additionally, $R(m)$ can be redefined as Equation (4) [10].

$$R(m) = \int_{-\pi}^{\pi} S(\omega)cos\omega m, d\omega \tag{4}$$

Finally, utilizing the FFT and IFFT functions the autocorrelation of the window can be efficiently calculated using Equation (5) [9, 10].

$$R(m) = \frac{1}{N} IFFT\left(|FFT\left(x(n)\right)|^2\right) \tag{5}$$

The fundamental frequency, $F0$, of the resulting signal will be represented by the lag location with the greatest amplitude. For emotional speech, the lag is restricted to a range between 50 and 500Hz; this corresponds to the region of pitch perceivable by humans [2].

Four statistical features are extracted from the fundamental frequency analysis. This begins with the mean, range and variance of $F0$. Finally, the slope of the pitch is calculated and fundamental frequency slopes greater than 70Hz are filtered out.

The data capture and signal processing will continue for a user-definable period after the first segment is detected. In testing it was found that a 2-3 second window of processing was sufficient.

At this point the extracted features can be sent to one of two different communications units described in detail below.

2.3 Data Communication

There are two options for interacting with an NCS simulation. The choice depends on the role the speech system is playing in a particular investigation.

Path 1

The first communication path is a Java based socket interface from MATLAB to a NCSTools server [12]. NCSTools is a C++ based software system that provides a number of mechanisms for communicating with a running NCS simulation. NCSTools accepts plain text strings from any number of clients connected through the built-in socket server. Through a custom configuration file, users can assign these strings to input stimulus to, or simulation controls of, a running NCS instance. Similarly, NCSTools can be configured to process simulation reports in a number of different ways. The results of which can be sent to

Table 1: Incorrectly classified recordings from the Berlin Database [11].

Code	Speaker Info	German Phrase	English Translation	Emotion	Prediction
03b10Ab	male, 31 years	*Die wird auf dem Platz sein, wo wir sie immer hinlegen*	*It will be in the place where we always store it*	Fear	**Anger**
03a02Wb	male, 31 years	*Das will sie am Mittwoch abgeben*	*She will hand it in on Wednesday*	Anger	**Fear**
15b01Wc	male, 25 years	*Was sind denn das fr Tten, die da unter dem Tisch stehen*	*What about the bags standing there under the table*	Anger	**Fear**

connected clients through the server interface. This allows designers of remote tools to interface with a neural-simulation in a way that abstracts them from the details of the model. Thus providing a mechanism of reuse without modification for different models; only the NCSTools configuration needs to be changed.

The use of this path was intended for, but certainly not limited to, coupling with the statistical classification output of the ESP system. As the data is categorized the results can be sent to NCSTools. NCSTools can then activate defined brain regions or dynamically modify Hebbian learning rules. This paradigm provides a means for verbally interacting with a neurorobotic avatar. Additionally, the classified verbal responses can be used for rewarding correct behaviors and discouraging those considered incorrect.

Path 2

The second communication option is a direct socket interface to a running NCS simulation. Similar to the direct NCSTools option above, this is comprised of a Java client interface that connects to the C++ server implementation. This option facilitates future feature classification methods employing biologically realistic spiking neural models; a combination that has significant potential for researchers of both computational neuroscience and speech recognition.

3 Testing

To demonstrate the system's capabilities a series of tests were completed. The real-time capture and processing was verified using MATLAB's built-in timing tools. Additionally, MATLAB's Data Acquisition Toolbox provided mechanisms for reporting when the sampled data was not removed before the next sample was ready. Tests were performed on both the classified result (Path 1) and direct data connection (Path 2).

3.1 English Recordings

Some initial analysis was completed with non-trained participants speaking the emotionally ambiguous phrase, "Look what you've done," while expressing two basic emotions, Joy and Disappointment. These recordings were not analyzed by trained reviewers, so the emotion was categorized by the researcher team only. These tests, however, did provide an excellent test facility for development of the system.

3.2 Berlin Emotional Speech Database

Finally, to show how this system can perform on standardized data, a portion of the Berlin Database of Emotional Speech was used [11]. Although in German, EmoDB provides an excellent reference for testing emotional speech algorithms. In emotional speech research there is a lack of freely available databases, especially in English. It was for this reason that EmoDB was utilized.

There is considerable evidence that the classification of the extracted features described above are dependent on the gender of the speaker [2, 13, 14]. This motivated the use of only male recordings for initial testing purposes. Similarly, the performance of many emotional speech recognition systems show a strong speaker dependence [2, 13]. In this project however, it was decided multiple speakers would be allowed but the range of emotions was limited to Anger and Fear. Generally as more emotions are added to the classification system the accuracy will decrease [8, 15, 16]. However, Wu *et al.* [17], accomplished higher recognition rates of 7 of the Berlin Database emotions (disgust was excluded) using long-term spectro-temporal features. Similar results were accomplished by Vlasenko *et al.*, [18], using a combination Gaussian Mixture and Support Vector Machine. The computational cost of these methods would require further investigation for inclusion in real-time system similar to that proposed here. The purpose was not to demonstrate superiority over existing systems but to merely illustrate a classification scheme that can perform accurately in real-time.

Five recordings from both the Anger and Fear sets were randomly selected. The remaining recordings were then analyzed using the emotional speech recognition system. It should be noted here that the recordings were used directly and not sampled by the data acquisition toolbox. As the results of the first

tests demonstrated, the system does operate in real-time. After the speech features were extracted, a classification and regression tree was constructed in MATLAB based on the results. Having been successfully employed by other emotional speech processors [19], classification trees can be simple to construct and allow the inclusion of other features not utilized here (e.g. linguistic aspects of the speech signal).

Each segment, as detected by the system, was classified and used as training data in the tree construction. The 10 recordings were then analyzed using the classification tree. Unlike the two to three second analysis period activated by the normal system, the entire recording was used. Each segment was classified, with the result stored locally. When the recording was complete, the system would select the dominant emotion in the recording and send the result to the NCSTools server.

4 Results

The results for the English recordings are included here as an illustration of the concepts presented above. In Figure 2 the black lines represent the raw data recorded during the training session. The red lines are the speech energy of the signal as calculated by Equation (1) and the gray lines frame the automatically calculated segments selected for analysis. The extracted features are given in Figure 3. The red marks are from recordings classified as Joy with the blue representing Disappointment.

Test results with the Berlin Database were promising. Of the 10 randomly selected recordings, 3 were classified incorrectly. With the ESP system correctly identifying 3 out 5 recordings labeled as angry and 4 out 5 as fear correctly. This is summarized in the confusion matrix labeled Table 2. Although only two emotions were trained and classified, this result still illustrates the potential of the design.

Table 2: Results of Berlin Database [11] Testing.

		Actual	
		Fear	**Anger**
Predicted	**Fear**	4	2
	Anger	1	3

5 Conclusion

A unique real-time emotional speech recognition system was presented with some promising test results. As a proof-of-concept, this project's results are encouraging and have provided evidence that future work and extended applications will be possible.

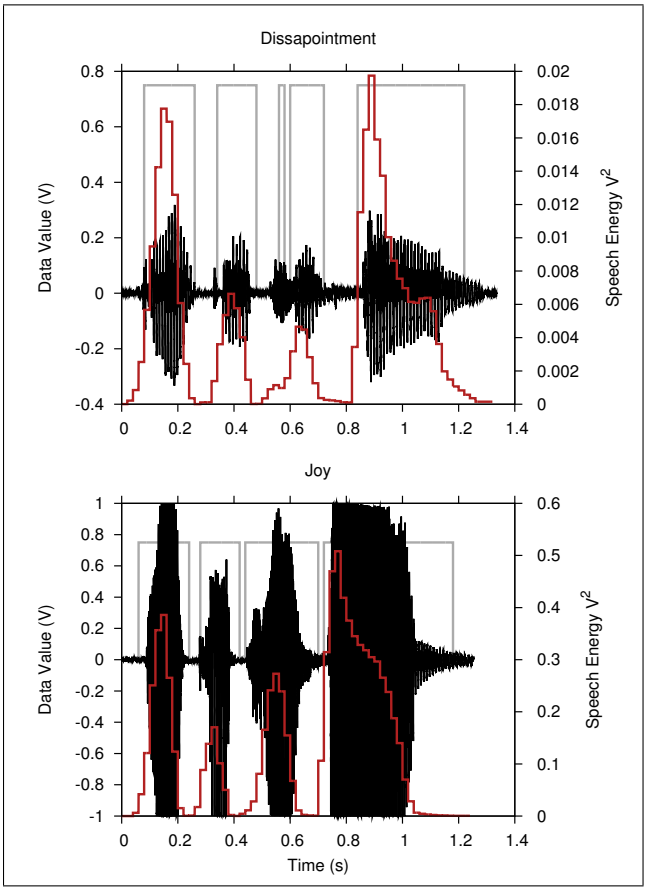

Figure 2: Results of male participant speaking the phrase "Look what you've done" with the acted emotions, joy and disappointed. Raw data is plotted with black lines. The Speech Energy is plotted in Red. The selected segments are framed by the gray Lines.

5.1 Applications

Beyond the applications to computational neuroscience and neurorobotics discussed previously, several applications for emotional processing have been identified by other researchers. These include call center monitoring [8, 20, 21], Human-Computer Interaction [16], aircraft pilot monitoring [2], and as a therapist diagnostics tool [2]. A possible addition could be applications in law enforcement. The ability to analyze the emotional state of both officers and civilians could provide law enforcement agents with a tool useful in both investigations and stressful situations.

5.2 Future Work

As this project progresses from a proof-of-concept to a functional research tool, there are several additions that need to be considered. Some of the more successful techniques referenced here will be considered as replacements for the current algorithms. In ad-

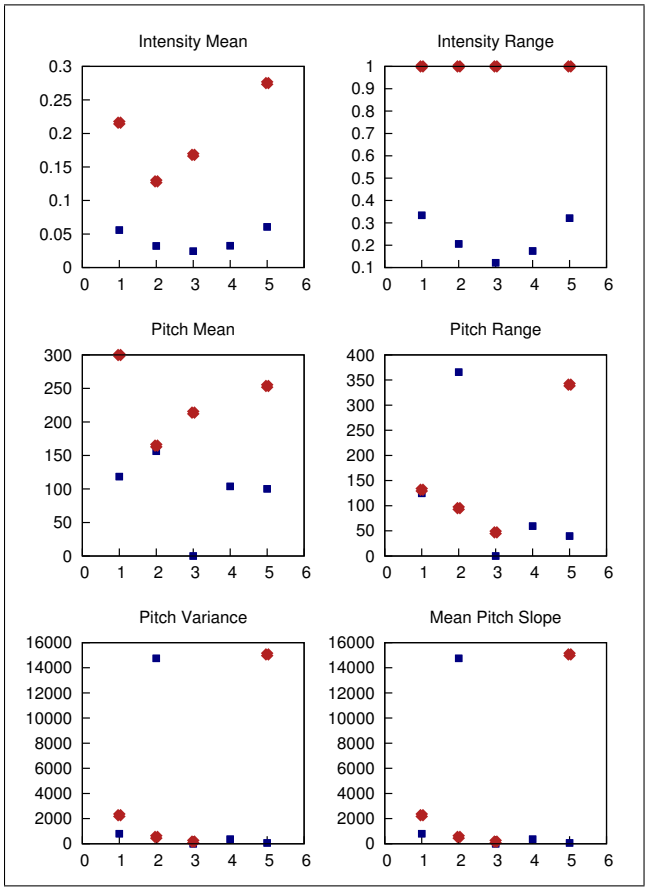

Figure 3: Extracted features of male participant. Values from recording classified as Joy are marked in red. The Disappointment values are marked in blue. The x-axis corresponds to the utterance as outlined in Figure 2. The intensity values are in volts and pitch values are in Hz.

dition, novel classification and segmentation concepts must be explored and integrated into the ESP system.

With more comprehensive algorithms the computational cost will inevitably increase. This will eventually lead to a loss of real-time performance and the need to explore new hardware and software platforms.

Acknowledgements

This work was supported in part by grants from the U.S. Defense Advanced Research Projects Agency (HR001109C001) and the U.S. Office of Naval Research (N000140110014).

References

[1] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389 – 405, 2005. Emotion and Brain.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, SEP 2006.

[3] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613 – 625, 2010.

[4] R. Brette, M. Rudolph, T. Carnevale, M. Hines, D. Beeman, J. M. Bower, M. Diesmann, A. Morrison, P. H. Goodman, F. C. Harris, Jr., M. Zirpe, T. Natschlager, D. Pecevski, B. Ermentrout, M. Djurfeldt, A. Lansner, O. Rochel, T. Vieville, E. Muller, A. P. Davison, S. El Boustani, and A. Destexhe, "Simulation of networks of spiking neurons: A review of tools and strategies," *Journal of Computational Neuroscience*, pp. 349–398, Nov. 2006.

[5] P. H. Goodman, S. Buntha, Q. Zou, and S.-M. Dascalu, "Virtual neurorobotics (vnr) to accelerate development of plausible neuromorphic brain architectures," *Frontiers in Neurorobotics*, 2007.

[6] P. H. Goodman, Q. Zou, and S.-M. Dascalu, "Framework and implications of virtual neurorobotics," *Frontiers in Neuroscience*, p. 5, 2008.

[7] T. Vogt and E. Andr, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *in ICME05*, 2005.

[8] V. A. Petrushin, "Emotion in speech: Recognition and application to call centers," in *In Engr*, pp. 7–10, 1999.

[9] S. J. Orfanidis, *Optimum Signal Processing An Introduction*. MacMillian Publishing Company, 1985.

[10] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. Marcel Decker, Inc., 2001.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech 2005*, pp. 1517–1520, September 2005.

[12] C. Thibeault, F. Harris, and P. Goodman, "Breaking the virtual barrier: real-time interactions with spiking neural models," *BMC Neuroscience*, vol. 11, no. Suppl 1, p. P73, 2010.

[13] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, pp. 763–786, 2007.

[14] T. Vogt and E. Andr, "Improving automatic emotion recognition from speech via gender differentiation," in *In Proc. Language Resources and Evaluation Conference (LREC 2006)*, 2006.

[15] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in *In SpeechEmotion-2000*, pp. 207–212, 2000.

[16] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved emotion recognition with a novel speaker-independent feature," *Mechatronics, IEEE/ASME Transactions on*, vol. 14, pp. 317 –325, June 2009.

[17] S. Wu, T. Falk, and W.-Y. Chan, "Automatic recognition of speech emotion using long-term spectro-temporal features," in *Digital Signal Processing, 2009 16th International Conference on*, pp. 1 –6, 2009.

[18] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," in *INTERSPEECH-2007*, pp. 2225–2228, 2007.

[19] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1145 –1154, july 2006.

[20] C. M. Lee and S. S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 293–303, 2005.

[21] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. ICSLP 2002*, pp. 2037–2040, 2002.