ISSN 2319-8885,Volume01,Issue No. 03

International Journal of Scientific Engineering and Technology Research

www.semargroups.org

Jul-Dec 2012, P.P. 186-191

Analysis of Speech Emotion Detection using Kullback Leibler Divergence Based on MFCC and Vector Quantization

T.KRANTHI KUMAR[1], K.SREEVANI[2] ,G.SAI KUMAR[3]

[1]M.Tech, ECE Dept, Bharat Institute of Engineering and Technology,AP-India,e-mail: kranthi0510@gmail.com,
[2]M.Tech, ECE Dept, Indur Institute of Engineering and Technology,AP-India,e-mail: skonamoni@gmail.com,
[3]M.Tech,ECE Dept,Vardhaman College of Engineering and Technology,AP-India,e-mail:saikumar418@gmail.com.

Abstract: In this project, is to convert the speech waveform, using digital signal processing tools, to a set of features at a considerably lower information rate for further analysis, the voice is enriched to convey not only the intended semantic message but also the emotional state of the speaker. The pitch contour is one of the important properties of speech that is affected by this emotional modulation. Although pitch features have been commonly used to recognize emotions, it is not clear what aspects of the pitch contour are the most emotionally salient. This paper presents an analysis of the statistics derived from the pitch contour. First, pitch features derived from emotional speech samples are compared with the ones derived from neutral speech, by using symmetric Kullback–Leibler distance. the speech can be modeled by some speech parameters like Mel frequency Cepstrum coefficients (MFCC) are the most widely used parameters in area of speech processing. We also have employed the same in our research. LPCs are derived on the assumption that speech signal is linear in nature and MFCCs are derived on the assumption that signal is logarithmic in nature this emotional modulation.

Keyword: MFCC, Emotional speech analysis, emotional speech recognition, pitch contour analysis.

1. INTRODUCTION

EMOTION plays a crucial role in day-to-day interpersonal human interactions. Recent findings have suggested that emotion is integral to our rational and intelligent decisions. It helps us to relate with each other by expressing our feelings and providing feedback. This important aspect of human interaction needs to be considered in the design of human–machine interfaces (HMIs) [1]. To build interfaces that are more in tune with the users' needs and preferences, it is essential to study how emotion modulates and enhances the verbal and nonverbal channels in human communication. Speech prosody is one of the important communicative channels that is influenced by and enriched with emotional modulation. The intonation, tone, timing, and energy of speech are all jointly influenced in a nontrivial manner to express the emotional message [2]. The standard approach in current emotion recognition systems is to compute high-level statistical information from prosodic features at the sentence-level such as mean, range, variance, maximum, and minimum of F0 and energy.

The speech signal is a slowly timed varying signal (it is called quasi-stationary). An example of speech signal is shown in Figure 1. When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Spectrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and will be described in this paper.

MFCC are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The process of computing MFCC is described in more detail next.

This paper focuses on one aspect of expressive speech prosody: the F0 (pitch) contour. The goal of this paper is twofold. The first is to study which aspects of the pitch contour are manipulated during expressive speech (e.g., curvature, contour, shape, dynamics). For this purpose, we present a novel framework based on Kullback–Leibler divergence (KLD) and logistic regression models to identify, quantify, and rank the most emotionally salient aspects of the F0 contour. Different acted emotional databases are used for the study, spanning different speakers, emotional categories and languages (English and German). First, the symmetric Kullback–Leibler distance is used to compare the distributionsof different pitch statistics (e.g., mean, maximum) between emotional speech and reference neutral speech. Then, a logistic regression analysis is implemented to discriminate emotional speech from neutral speech using the pitch statistics as input.

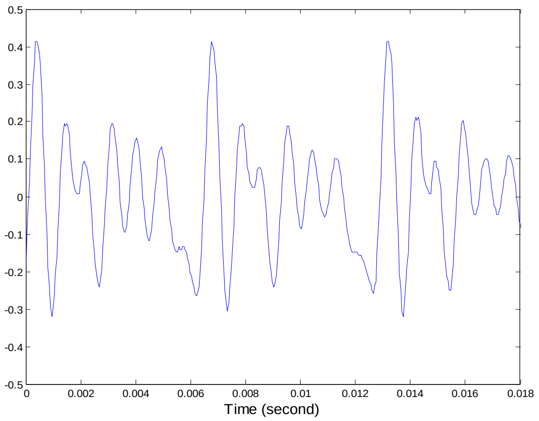

Figure1. Example of speech signal

These experiments provide insights about the aspects of pitch that are modulated to convey emotional goals. The second goal is to use these emotionally salient features to build robust prosody speech models to detect emotional speech. In our recent work, we introduced the idea of building neutral speech models to discriminate emotional speech from neutral speech [6]. This approach is appealing since many neutral speech corpora are available, compared to emotional speech corpora, allowing the construction of robust neutral speech models. Furthermore, since these models are independent of the specific emotional databases, they can be more easily generalized to real-life applications [7]. While the focus on our previous paper was on spectral speech models, this paper focuses on features derived from the F0 contour. Gaussian mixture models (GMMs) are trained using the most discriminative aspects of the pitch contour, following the analysis results presented in this paper.

The results reveal that features that describe the global aspects (or properties) of the pitch contour, such as the mean, maximum, minimum, and range, are more emotionally salient than features that describe the pitch shape itself (e.g., slope, curvature, and inflexion). However, features such as pitch curvature provide complementary information that is useful for emotion discrimination. The classification results also indicate that the models trained with the statistics derived over the entire sentence have better performance in terms of accuracy and robustness than when they are trained with features estimated over shorter speech regions (e.g., voiced segments).

II. PROPOSED METHODOLOGY

The fundamental frequency or F0 contour (pitch), which is a prosodic feature, provides the tonal and rhythmic properties of the speech. It predominantly describes the speech source rather than the vocal tract properties. Although it is also used to emphasize linguistic goals conveyed in speech, it is largely independent of the specific lexical content of what is spoken in most languages [7]. The fundamental frequency is also a supra-segmental speech feature, where information is conveyed over longer time scales than other segmental speech correlates such as spectral envelope features. Therefore, rather than using the pitch value itself, it is commonly accepted to estimate global statistics of the pitch contour over an entire utterance or sentence (sentence-level) such as the mean, maximum, and standard deviation.

However, it is not clear that estimating global statistics from the pitch contour will provide local information of the emotional modulation [9]. Therefore, in addition to sentence-level analysis, we investigate alternative time units for the F0 contour analysis. Examples of time units that have been proposed to model or analyze the pitch contour include those at the foot-level [8], word-level [10], and even syllable-level [1]. In this paper, we propose to study the pitch features extracted over voiced regions (hereon referred as voiced-level). In this approach, the frames are labeled as voiced or unvoiced frames according to their F0 value (greater or equal to zero). Consecutive voiced frames are joined to form a voiced region over which the pitch statistics are estimated. The average duration of this time unit is 167 ms. The lower and upper quartiles are 60 and 230 ms, respectively. The motivation behind using voiced region as a time unit is that the voicing process, which is influenced by the emotional modulation, directly determines voiced and unvoiced regions. Therefore, analysis along this level may shed further insights into emotional influence on the F0 contour not evident from the sentence level analyses.

From a practical viewpoint, voiced regions are easier to segment compared to other short time units, which require forced alignment (word and syllable) or syllable stress detections (foot). In real-time applications, in which the audio is continuously recorded, this approach has the advantage that smaller buffers are required to process the audio. Also, it does not require pre-segmenting the input speech into utterances. Both sentence- and voiced-level pitch features are analyzed in this paper.

For the sake of generalization, the results presented in this paper are based on four different acted emotional databases (three for training and testing and one for validation) recorded from different research groups and spanning different emotional categories. Therefore, some degree of variability in the recording settings and the emotional elicitation is included in the analysis. Instead of studying the pitch contour in terms of emotional categories, the analysis is simplified to a binary problem in which emotional speech is contrasted with neutral speech (i.e., neutral versus emotional speech). This approach has the advantage of being independent of the emotional descriptors (emotional categories or attributes), and it is useful for many practical applications such as automatic expressive speech mining. In fact, it can be used as a first step in a more sophisticated multiclass emotion recognition system in which a second level classification would be used to achieve a finer emotional description of the speech.

III. Mel-frequency Spectrum coefficients processor

A block diagram of the structure of an MFCC processor is given in Figure 3. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC are shown to be less susceptible to mentioned variations.

Frame Blocking: In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples and so on. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

Windowing: The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1 \qquad \textbf{(1)}$$

Typically the Hamming window is used, which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \quad \textbf{(2)}$$

Fast Fourier Transform (FFT): The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {x_n}, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,..., N-1 \qquad \textbf{(3)}$$

In general X_k's are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence {X_k} is interpreted as follow: positive frequencies $0 \le f < F_s/2$ correspond to values $0 \le n \le N/2-1$, while negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, F_s denotes the sampling frequency. The result after this step is often referred to as spectrum or periodogram.

Mel-frequency Wrapping: As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

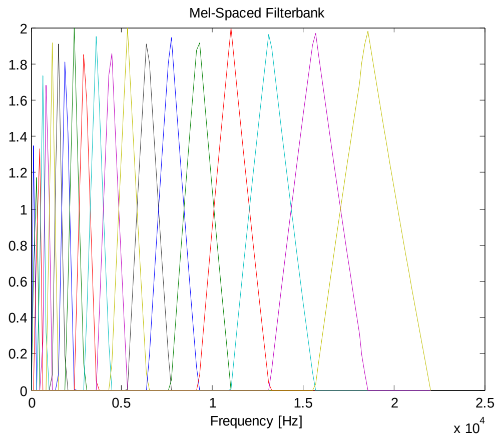

Figure2. An example of Mel-spaced Filter bank

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the Mel-scale (see Figure 2). That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel frequency interval. The number of Mel spectrum coefficients, K, is typically chosen as 20. Note that this filter bank is applied in the frequency domain, thus it simply amounts to applying the triangle-shape windows as in the Figure 4 to the spectrum. A useful way of thinking about this Mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

Cepstrum: In this final step, we convert the log Mel spectrum back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those Mel power spectrum coefficients that are the result of the procedure described above, for each speech frame of around 30msec with overlap, a set of Mel-frequency cepstrum coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a Mel-frequency scale. This set of coefficients is called an acoustic vector. Therefore each input utterance is transformed into a sequence of acoustic vectors. In the next section we will see how those acoustic vectors can be used to represent and recognize the voice characteristic of the speaker. Furthermore, if there exists some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm.

The state-of-the-art in feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this project, the VQ approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook.

The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance.

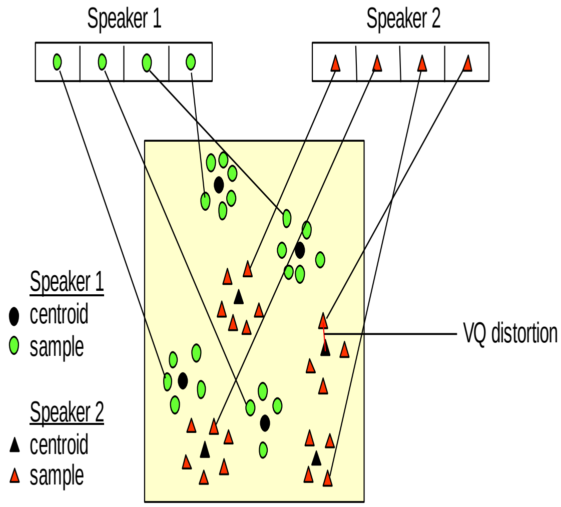

Figure 3. Conceptual diagram illustrating vector quantization codebook formation.

One speaker can be discriminated from another based of the location of centroids. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook.

IV. SIMULATION RESULTS

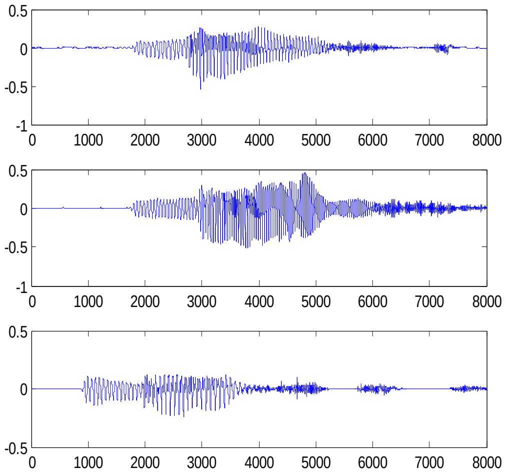

Figure4: Input Speech signal at different Emotions

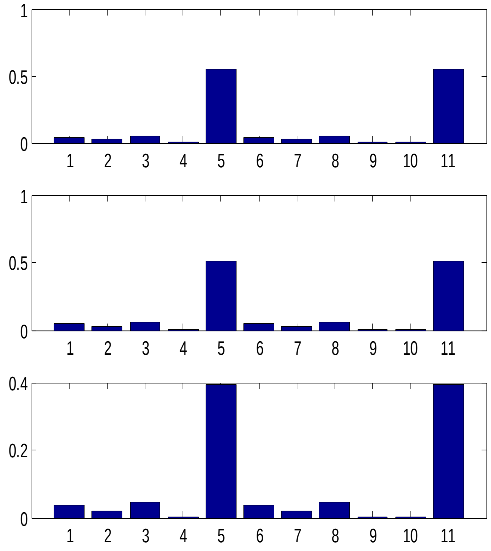

Figure5: Most emotional prominent features according to the average symmetric KLD ratio between features derived from emotional and neutral speech. The figures show the sentence-level (top) and voiced-level (bottom) features.

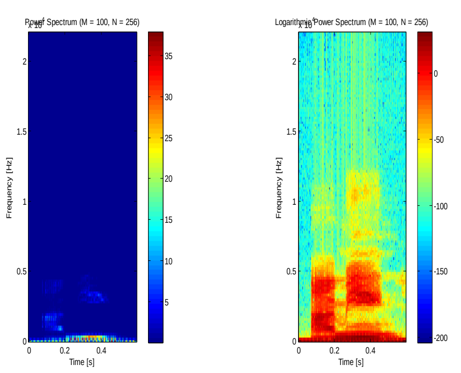

Figure6: Speech Signal Power Spectrum and Logarithmic Power Spectrum

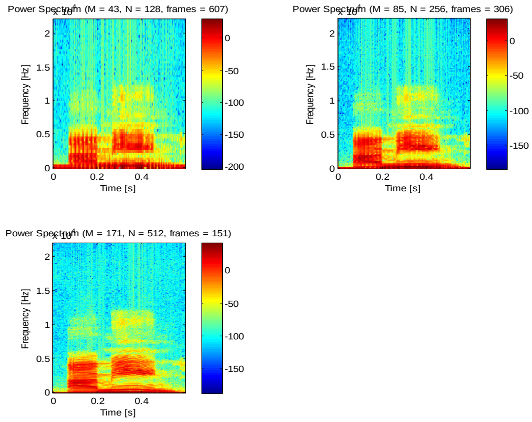

Figure7: Power Spectrum at different Emotions Salient Aspects

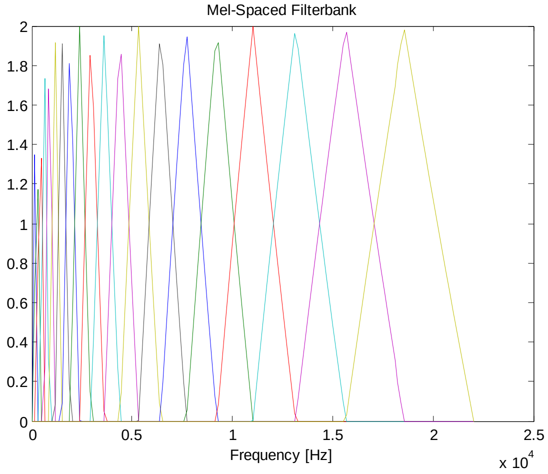

Figure8: MFCC filter bank.

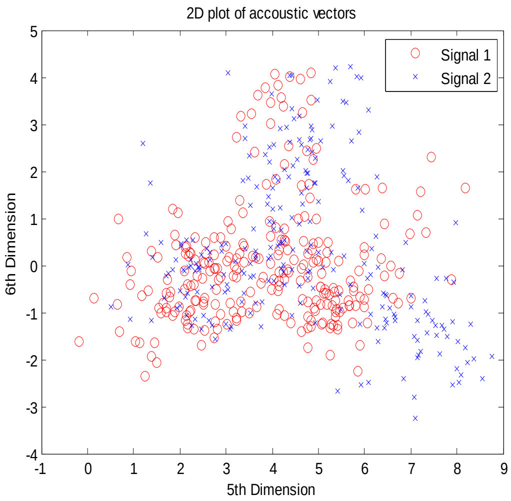

Figure9: Different Voice Emotion Estimation Comparison using Vector Quantization.

V. CONCLUSION

This paper presented an analysis of different expressive pitch contour statistics with the goal of finding the emotionally salient aspects of the F0 contour (pitch). For this purpose, two experiments were proposed. In the first experiment, the distribution of different pitch features was compared with the distribution of the features derived from neutral speech using the symmetric KLD with MFCC and Vector Quantization Method, Both experiments indicate that dynamic statistics such as mean, maximum, minimum, and range of the pitch are the most salient aspects of expressive pitch contour. The statistics were computed at sentence and voiced region levels. The results indicate that the system based on sentence-level features outperforms the one with voiced-level statistics both in accuracy and robustness, which facilitates a turn-by-turn processing in emotion detection.

VI. REFERENCES

[1] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1999.

[2] L.R Rabiner and R.W. Schafer, Digital Processing of peech Signals, Prentice-Hall, Englewood Cliffs, N.J., 1978.

[3] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, Signal Processing, Vol. ASSP-28, No. 4, August 1980.

[4] Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

[5] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol. ASSP-34, No. 1, pp. 52-59, February 1986.

[6] S. Furui, "An overview of speaker recognition technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[7] R. W. Picard, "Affective Computing," MIT Media Laboratory Perceptual Computing Section, Cambridge, MA, USA, Tech. Rep. 321, Nov. 1995.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human–computer interaction," IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan.2001.

[9] A. Álvarez, I. Cearreta, J. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay, "Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken spanish and standard basque language," in Proc. 9th Int. Conf. Text, Speech and Dialogue (TSD 2006), Brno, Czech Republic, Sep. 2006, pp. 565–572.

[10] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in Proc. XIV Eur. Signal Process. Conf. (EUSIPCO'06), Florence, Italy, Sep. 2006, pp. 929–932.

[11] M. Sedaaghi, C. Kotropoulos, and D. Ververidis, "Using adaptive genetic algorithms to improve speech emotion recognition," in Proc. Int. Workshop Multimedia Signal Process. (MMSP'07), Chania, Crete, Greece, Oct. 2007, pp. 461–464.

[12] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in Proc. Interspeech'07—Eurospeech,Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

[13] E. Douglas-Cowie, L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in Proc. 9th Eur. Conf. Speech Commun. Technol. (Interspeech'05), Lisbon, Portugal, Sep. 2005, pp. 813–816.

[14] P. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?," Psychol. Bull., vol. 129, no. 5, pp. 770–814, Sep. 2003.