Vocal Emotion Recognition with Log-Gabor Filters

Yu Gu
Tilburg center for Cognition and Communication, School of Humanities, University of Tilburg
P.O.Box 90153
Tilburg, Netherlands
y.gu_1@uvt.nl

Eric Postma
Tilburg center for Cognition and Communication, School of Humanities, University of Tilburg
P.O.Box 90153
Tilburg, Netherlands
E.O.Postma@uvt.nl

Hai-Xiang Lin
Delft Institute of Applied Mathematics, Delft University of Technology
P.O.Box 5031
Delft, The Netherlands
h.x.lin@tudelft.nl

ABSTRACT

Vocal emotion recognition aims to identify the emotional states of speakers by analyzing their speech signal. This paper builds on the work of Ezzat, Bouvrie and Poggio [5] by performing a spectro-temporal analysis of affective vocalizations by decomposing the associated spectrogram with 2D Gabor filters. Based on the previous studies of the emotion expression in voices and the turn out display in spectrogram, we assumed that each vocal emotion has a unique spectro-temporal signature in terms of orientated energy bands which can be detected by properly tuned Gabor filters. We compared the emotion-recognition performances of tuned log-Gabor filters with standard acoustic features. The experimental results show that applying pairs of log-Gabor filters to extract features from the spectrogram yields a performance that matches the performance of an approach based on traditional acoustic features. Their combined emotion recognition performance outperforms state-of-the-art vocal emotion recognition algorithms. This leads us to conclude that tuned log-Gabor filters support the automatic recognition of emotions from speech and may be beneficial to other speech-related tasks.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Affective Computing, Speech, Emotion Recognition, Log-Gabor Filter

1. INTRODUCTION

Emotion recognition from speech plays a significant role in human-machine interaction, which is becoming increasingly important given the immersion of computational devices in daily life [13]. Apart from verbal signals, human speech contains non-verbal signals that provide (amongst others) information on the affective state of the speaker, such as, intensity, pitch and other spectral features. Automatic emotion recognition from speech contributes to the interaction of humans and machines, because it allows algorithms to detect nonverbal cues of users about, for instance, their level of frustration or stress.

AVEC'15 October 26 2015 Brisbane, Australia.

ACM ACM 978-1-4503-3743-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808196.2811635.

1.1 Features for Automatic Emotion Recognition

Current automatic emotion recognition systems rely on machine learning. In these systems, features that are assumed to be of relevance to the classification task at hand, are extracted from the speech signal. Classifiers are trained on the feature representations of the speech signal to estimate the appropriate classes for recognition. The construction of the appropriate features, so-called *feature construction* is crucial to the recognition performance. Traditional speech processing and recognition methods often rely on temporal and spectral features that have proven relevant for speech-related tasks [3]. Well-known examples of such established features are Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) which consist of a wide variety of measurements of the speech signal.

Inclusion of all potentially relevant features ensures that all relevant measurements are present, but a large number of features gives rise to the curse of dimensionality which deteriorates generalization performance [8]. A viable alternative to feature construction is called feature learning, in which the relevant features are obtained automatically from the raw speech signals [11], but this approach demands considerable computational resources and requires extensive experimentation to find the appropriate parameters. If domain knowledge is available to guide the selection of features, feature construction may be feasible.

This paper employs feature construction to recognize emotions from speech. To avoid the curse of dimensionality, the number of features is kept small by using domain knowledge to guide the feature construction.

1.2 Treating the Spectrogram as an Image

Many speech analysis approaches rely on spectral information. The spectrogram is a widely used representation of spectral information for auditory signal analysis in a wide range of application domains, such as speech discrimina-

tion [4], environmental sound classification [16], automatic speech recognition [12], and investigation of the personality and likeability [2].

We treat the spectrogram as an image by performing analyses of its local spectro-temporal structure. The analyses are performed using standard image processing, i.e., two-dimensional Gabor filters which are locally tuned to the orientations of energy bands in the spectrogram. The idea of using 2D Gabor filters for analysing the spectrogram is due to Ezzat, Bouvrie and Poggio [5]. Their focus was on the detection of general speech-related patterns in the spectrogram. Our focus is on the extraction of affective speech. To determine the contribution of the visual features extracted by means of 2D Gabor filters, we will perform a comparative evaluation of the following four types of features in a recognition task: acoustic features (e.g. MFCC and LPCC features), (2) untuned Gabor filters, (3) tuned Gabor filters, (4) combination of (1) and (3).

1.3 Outline

The remainder of this paper is structured as follows. Section 2 outlines the 2D Gabor analysis of affective speech and specifies the different approaches used in our empirical study. In Section 3, the experimental set-up is detailed. The experiment results are presented in section 4. Finally, Section 5 discusses the results and draws conclusions.

2. VISUAL ANALYSIS OF THE SPECTROGRAM

The idea to treat the spectrogram as an image that can be analyzed using image-processing methods follows from an examination of an example of a spectrogram. Figure 1 illustrates a part of a speech spectrogram [19]. The horizontal axis represents the time and the vertical axis represents the frequency. Red color means a high energy value and blue means low energy value. For periodic vocal signals, the spectrogram contains parallel bands that correspond to the partials of the complex tone generated by the vocal chords. The inset shows an example of a periodic fragment with horizontal bands. The horizontal orientation of the energy bands reflects the constant frequency over the selected period of time. Properly-tuned 2D Gabor filters respond to the width (spatial frequency) and orientation of bands in the spectrogram.

Therefore, by convolving the spectrogram with a Gabor filter of a given spatial frequency and orientation, the convolved spectrogram represents spectro-temporal patterns with the associated spatial frequency (width) and orientation, respectively. In this paper, we only tune the orientation of the Gabor filters and average over a range of spatial frequencies encompassing the widths of the energy bands of interest.

2.1 Log-Gabor Filters

The original (1-dimensional) Gabor filter was proposed by Dennis Gabor [7] to deal with the inherent uncertainty in determining the temporal localization and frequency. Measuring the frequency of a signal requires a certain temporal extent over which to make the measurement. In one dimension, the Gabor filter corresponds to a sine wave weighted by a Gaussian envelope which combines localization (the mean of the Gaussian) with frequency determination (the frequency of the sinusoid). A limitation of the Gabor filter is that it can have a non-zero DC value (which the mean value of a wavevorm) for certain bandwidths. An improvement of the original Gabor filter proposed by Field [6], the log-Gabor filter ensures a zero DC value by defining the Gabor filter on a logarithmic frequency scale. In our approach, we adopt log-Gabor filters.

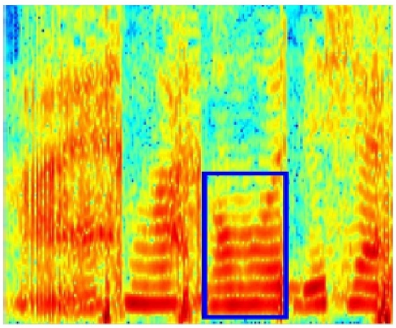


Figure 1: Spectrogram of the utterance "He is a good person" in Chinese expressed with a neutral emotion. The part enclosed by the blue rectangle corresponds to the fragment "a good person".

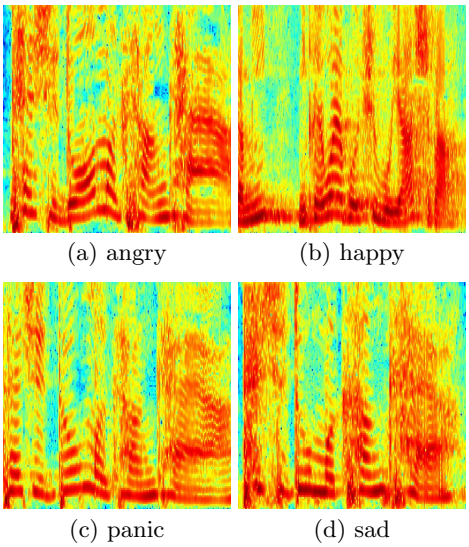
2.2 Visual Feature Extraction

To illustrate the application of log-Gabor filters to affective vocal expressions, Figure 2 shows four spectrograms of the utterance "He is a good person" in Chinese. The utterances differ in their emotion. Figure 2 shows the spectrograms for the four emotional expressions: (a) *angry*, (b) *happy*, (c) *panic* and (d) *sad*. By comparing the four spectrograms with the neutral spectrogram shown in Figure 1, their differences become apparent. Whereas for the neutral emotion the energy bands are mainly horizontal, for the four emotions shown in Figure 2, different orientation patterns are present. The differences in orientations reflect the spectro-temporal dynamics of vocal pitch which could rise over time (upward orientation), remain stable (horizontal), or fall over time (downward orientation). Log-Gabor filters tuned to the appropriate orientations may help to extract these subtle differences from the spectrograms.

2.3 Tuned Gabor Filters

Previous studies contributed to the understanding of how different types of emotions are vocally expressed. [10] [1]. Hammerschmidt and Jurgens found the spectro-temporal energy bands useful for describing acoustic characteristics of emotional vocal expressions. [9]. They observed that different vocal emotions were associated with different energy bands. They evaluated five types of vocal emotions: anger, happy, panic, sadness and the neutral.

Figure 3 displays an example of an angry utterance "So bad" in Chinese. The spectro-temporal representation consists of two segments. The left segment consists of parallel energy bands that move upwards. The right segment contains parallel energy bands that move downwards. Hammerschmidt and Jürgens [9] measured for the bands within each segment the minimum and maximum frequency values. In the figure, these extreme values are represented by squares (minimum values) and circles (maximum values). The slope



(a) angry (b) happy

(c) panic (d) sad

Figure 2: Four spectrograms of the utterance "He is a good person" in Chinese each spoken with a different emotion.

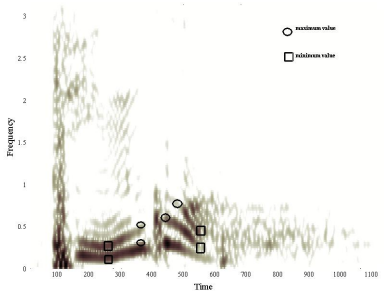


Figure 3: Spectrogram of the phrase "So bad" in Chinese expressed with an angry vocal emotion. The energy bands have an with a upward and sharp downward contour orientation. The minimum and maximum values of an energy band are indicated by a square and circle, respectively.

of the line connecting the minimum and maximum value quantifies the orientation of the energy bands..

We translated the quantitative orientation measurements of Hammerschmidt and Jürgens into five qualitative descriptions: horizontal, fast upward, slow upward, fast downward, slow downward. Table 1 specifies the descriptions for each of the five types of emotions.

To detect the vocal emotions from their spectro-temporal signature, we defined two sets of tuned Gabor filters. As discussed above, the angle could be positive or negative. The orientation of the filter was set as follows: the horizontal is 0 degree, fast upward is 45 degree, slow upward is 30. The downward slopes were defined by negative angles. No attempt was made to optimize the orientation through machine learning. We experimented with single filters (covering one segment) and double filters (covering two neighboring segments). This first set consisted of single filters tuned to the dominant orientation in the associated spectrogram.

Table 1: Qualitative descriptions of the slopes of the first and second segment of five vocal emotions.

Emotion	First Segment	Second Segment
Neutral	Horizontal	Horizontal
Angry	Fast upward	Slow downward
Happy	Fast upward	Fast downward
Panic	Slow upward	Fast downward
Sad	Slow downward	Slow downward

Table 2 lists the orientations of the single log-Gabor filters designed to detect the five emotions (including neutral).

Table 2: Specification of the single log-Gabor filters tuned to the five emotions.

Vocal emotion	Gabor filter	orientation
Neutral	$G^1_{neutral}$	0°
Angry	G^1_{angry}	45°
Happy	G^1_{happy}	45°
Panic	G^1_{panic}	30°
Sad	G^1_{sad}	−30°

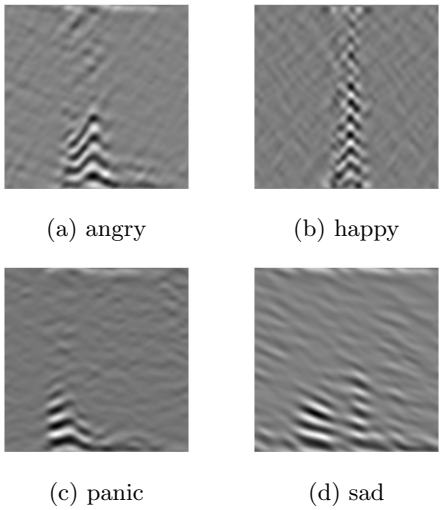
The second set of tuned Gabor filters consisted of horizontally contiguous pairs of filters that are tuned to the dominant combinations of orientations in the spectrograms. Also in this case, the combinations were estimated from a representative sample of emotional expressions. Table 3 lists the orientations of the log-Gabor filter pairs designed to detect the emotions. The orientations of the left and right filters of each contiguous pair are specified in the columns labelled left and right, respectively. Neutral and Sad only have one dominant orientation direction. Anger, Happy, and Panic have two different orientations. Figure 4 displays the Gabor filter pair tuned to detect the characteristic orientations of the spectrogram associated with panic.

Table 3: Specification of the log-Gabor filter pairs tuned to the five emotions.

Vocal emotion	Gabor filter	left	right
Neutral	$G^2_{neutral}$	0°	0°
Angry	G^2_{angry}	45°	−30°
Happy	G^2_{happy}	45°	−45°
Panic	G^2_{panic}	30°	−45°
Sad	G^2_{sad}	−30°	−30°

Figure 5 illustrates the result of convolving the four emotional expressions happy, angry, panic, and sad with the filter pairs G^2_{angry}, G^2_{happy}, G^2_{panic} and G^2_{sad}, respectively. By comparing the four convolved images, the specific orientation patterns of the Gabor pairs are clearly visible. Provided that the tuned filters respond selectively to the emotion-specific orientations in the spectrograms, they support the automatic recognition of emotions from speech.

Figure 4: Illustration of G^2_{panic}.



(a) angry (b) happy

(c) panic (d) sad

Figure 5: Convolution images obtained by convolving the spectrograms in Figure 2 with with the associated Gabor filter pairs listed in Table 3.

3. EXPERIMENTAL SET-UP

To determine the contribution of tuned Gabor filters to the automatic recognition of affective speech, a comparative evaluation was performed on acoustic features, untuned Gabor filters (all orientations), tuned Gabor filters (single orientations and pairs of orientations), and the combination of acoustic features with tuned Gabor filters.

3.1 Affective Speech Corpus

The performances of the automatic emotion recognition with different features was evaluated on the Mandarin Affective Speech (MAS) corpus (MAS, 2007) [18]. The MAS corpus contains utterance of 68 speakers (23 females), which comprise recordings of them uttering sentences with different emotions. In order to avoid exaggerated expressions of emotion, the developers of the corpus asked listeners to judge the naturalness of the utterances. Those records jugded to be unnatural were discarded from the corpus. All utterances were recorded with a sampling rate of 8 kHz at 16 bits. The corpus includes five types of emotion: angry, happy, neutral, panic and sad. For each emotion, speakers read 15 different sentences and every sentence is repeated four times. The total number of utterance is equal to 20.400 (number of emotions × number of sentences × number of repetitions × number of speakers). The corpus was obtained through the Linguistic Data consortium[1]. Table 1 summarizes the records in the corpus.

Table 4: Summary of the Mandarin Affective Speech corpus.

Mandarin Affective Speech corpus	
Number of emotions	5
Number of sentences	15
Number of repetitions	4
Number of speakers	68
Total number of utterances	20.400

3.2 Feature evaluation

The evaluation of the Gabor filters was based on the following four steps applied to each utterance in the corpus: (i) spectrogram calculation, (ii) convolution with Gabor filters, (iii) dimensionality reduction, and (iv) classification. In what follows, each of the four steps is specified in detail.

Spectrogram Calculation

Each auditory signal (utterance) was transformed into a spectrogram using Matlab's spectral analysis function employing the short-time Fourier transform with a 20 ms Hamming window and an overlap of half window length. The dimensions of the spectrograms were 512∗512 pixels.

Convolution with Gabor filters

Kovesi's log-Gabor functions [2] were used to perform the convolutions on the four quadrants of each spectrogram. The parameters used are as follows. Number of orientations = 12 in equal steps covering 360 degrees for "untuned". For "tuned" the 8 orientations and orientation pairs specified in Tables 2 and 3 were used. Number of scales = 12; Minimum wavelength = 3 and sigmaOnf = 0.8. For each of the four spectrogram region, each scale and orientation yields a single convolution image. The convolution values within each image is averaged yielding a total number of Gabor energy values equal to 4× number of orientations × number of scales.

Dimensionality Reduction

To reduce the redundancy of the Gabor energy values, Principal Component Analysis is applied. We used the PCA function incorporated in the dimensionality reduction toolbox[17] and optimized the number of retained components from 10 to 80 in steps of 10.

Classification

The dimensionality-reduced Gabor energy values were used for training an SVM classifier. We used the WEKA[3] implementation, with an RBF kernel. The parameter values *gamma* and C were optimized by means of grid search.

Acoustic Features

All the acoustic features used in our experiment correspond to the baseline features of the Interspeech challenge [15].

[1]http://catalog.ldc.upenn.edu/LDC2007S09/
[2]http://www.csse.uwa.edu.au/ pk/research/matlabfns/
[3]http://www.cs.waikato.ac.nz/ml/weka/

This is the current sate-of-the-art emotion recognition features that make the acoustic features convinced. The acoustic feature set for each utterance consisted of the following features: Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC), zero-crossing rate, speech rate, pitch, formants $(1-3)$, magnitude. Each feature was represented by four statistical descriptors: mean, maximum, minimum and standard deviation. We used the Voicebox[4] software for extracting the acoustic features.

In the comparative evaluation, these features are treated similarly as the Gabor features, i.e., application of PCA and classification by means of SVM. In the experiment involving acoustic features and Gabor features, all features are combined into a single feature vector and submitted to PCA and SVM.

3.3 Evaluation Procedure

The evaluation of the emotion recognition performance was determined for each set of features using cross-validation procedures. To avoid overfitting due to the PCA and the SVM parameter optimization, the evaluation was performed using both leaving-one-speaker and leaving-one-sentence out validations, in which optimization was performed in the outer each-fold leave-out cycle and the evaluation in the inner cycle.

4. RESULTS

Figure 6 shows box plots of the recognition performances obtained for the five sets of features: (a) acoustic features, (b) untuned Gabor filters, (c) tuned Gabor filters, 9d) tuned Gabor filter pairs, and (e) acoustic + tuned Gabor filter pairs. Traditional acoustic features outperform the untuned and tuned (single) Gabor filters. Apparently, encoding oriented energy bands in the spectrogram does not lead to a better performance than obtained with the traditional acoustic features. The improved performance of the tuned Gabor filters as compared to the untuned ones (which of course include the tuned orientations), indicates that reduction of the features is beneficial to the performance. Interestingly, the tuned Gabor filter pairs perform at a par with the acoustic features. Their combination yields the best performance overall, suggesting that they capture partly non-overlapping vocal characteristics.

Tables 5-9 show the confusion tables for the five features sets. Each confusion table shows in terms of percentage correct how often each emotion is recognized correctly (diagonal entries) or incorrectly (off-diagonal entries). These tables show that for all sets of features examined the emotion-specific performances agree quite well.

Table 5: Confusion table acoustic features.

	Performance (%)				
Emotion	Angry	Happy	Neutral	Panic	Sad
Angry	90.7%	3.8%	2.4%	2.0%	1.1%
Happy	4.2%	91.1%	1.5%	2.3%	0.9%
Neutral	2.3%	1.9%	92.6%	1.5%	1.7%
Panic	2.6%	1.9%	0.8%	92.2%	2.5%
Sad	1.5%	1.4%	3.2%	1.9%	91.9%

[4]www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html/

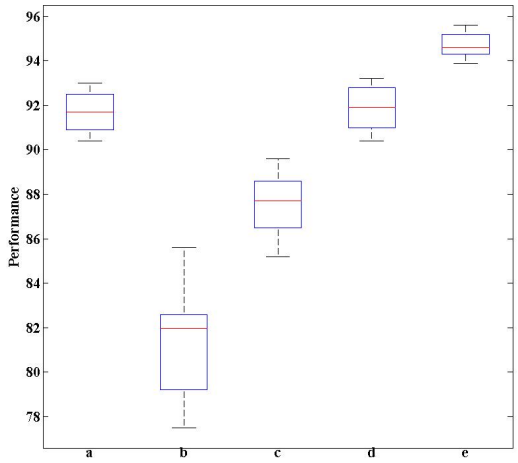


Figure 6: Recognition performances obtained for the five sets of features.

Table 6: Confusion table untuned Gabor.

	Performance (%)				
Emotion	Angry	Happy	Neutral	Panic	Sadness
Angry	80.9%	6.7%	3.5%	5.1%	3.8%
Happy	7.2%	79.1%	3.8%	5.7%	4.3%
Neutral	4.7%	3.9%	83.9%	3.5%	4.0%
Panic	4.8%	6.4%	2.9%	81.7%	4.2%
Sadness	4.4%	5.1%	3.7%	4.6%	82.2%

5. CONCLUSIONS

Our assessment of the use of Gabor filters to extract "visual" spectro-temporal features from the speech spectrogram revealed the feasibility of the idea put forward originally by Ezzat, Bouvrie and Poggio [5]. Especially the use of tuned Gabor filter pairs to perform a second-order analysis of the spectrogram led to good results. We believe that the reason for the good performance obtained is due to two reasons. The first reason is that we performed manual feature selection by inspecting many spectrograms of emotional speech. This leads to a limited set of features which avoids the curse of dimensionality that capture task-relevant information from the spectrogram. The second reason is the fact that Gabor filters detect relevant characteristics in the spectrogram for the task at hand. The patterns of orientations in the spectrogram reflect the time-varying frequency compositions of the emotional utterances. The orientation tuning of Gabor filters is highly suitable to encode these patterns.

We have ignored the role of spatial frequency (the width of the energy bands) by including a wide range of spatial frequencies in our tuned Gabor filters. Possibly, an improved tuning in terms of spatial frequency may further enhance the results obtained. This is left to future study.

Our findings lead us to conclude that tuned log-Gabor filters support the automatic recognition of emotions from speech and may be beneficial to other speech-related tasks.The experiments are carried out on a database with acted data,

Table 7: Confusion table tuned Gabor filters.

	Performance (%)				
Emotion	Angry	Happy	Neutral	Panic	Sadness
Angry	87.1%	4.5%	2.7%	3.4%	2.3%
Happy	4.0%	86.6%	2.4%	3.8%	3.1%
Neutral	3.3%	2.9%	89.4%	2.5%	1.9%
Panic	3.7%	4.6%	2.8%	85.7%	3.2%
Sad	3.2%	3.0%	2.9%	2.4%	88.5%

Table 8: Confusion table tuned Gabor filter pairs.

	Performance (%)				
Emotion	Angry	Happy	Neutral	Panic	Sadness
Angry	91.6%	2.7%	2.2%	1.9%	1.6%
Happy	2.9%	92.1%	1.1%	2.1%	1.8%
Neutral	1.9%	2.3%	93.2%	1.4%	1.2%
Panic	1.8%	2.9%	1.4%	91.9%	1.9%
Sadness	2.6%	2.2%	2.7%	1.8%	90.7%

it will be interesting to test the presented method on other speech databases in future work [14].

6. ACKNOWLEDGMENTS

Part of this work is funded by a China Scholarship Council (No.201206660009) awarded to the first author. The last author thanks the Timman Foundation for the partial financial support in this research.

7. REFERENCES

[1] T. Bänziger and K. R. Scherer. The role of intonation in emotional expressions. *Speech communication*, 46(3):252–267, 2005.

[2] H. Buisman and E. O. Postma. The log-gabor method: speech classification using spectrogram image analysis. In *INTERSPEECH*, 2012.

[3] L. Chen, X. Mao, Y. Xue, and L. Cheng. Speech emotion recognition: Features and classification models. *Digital Signal Processing*, pages 1154–1160, 2012.

[4] T. Chi, P. Ru, and S. A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[5] T. Ezzat, J. V. Bouvrie, and T. Poggio. Spectro-temporal analysis of speech using 2-d gabor filters. In *INTERSPEECH*, pages 506–509, 2007.

[6] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, 1987.

[7] D. Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[8] G. George and V. C. Raj. Review on feature selection techniques and the impact of svm for cancer classification using gene expression profile. *arXiv preprint arXiv:1109.1062*, 2011.

Table 9: Confusion table for the combination of acoustic features and tuned Gabor filter pairs.

	Performance (%)				
Emotion	Angry	Happy	Neutral	Panic	Sad
Angry	93.5%	2.4%	2.1%	1.7%	0.3%
Happy	3.2%	93.3%	1.0%	1.8%	0.7%
Neutral	1.5%	1.1%	96.5%	0.4%	0.5%
Panic	1.8%	1.3%	0.5%	94.3%	2.1%
Sad	0.5%	0.9%	2.7%	1.3%	94.6%

[9] K. Hammerschmidt and U. Jürgens. Acoustical correlates of affective prosody. *Journal of Voice*, 21(5):531–540, 2007.

[10] J. Kreiman and D. Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011.

[11] H. Lee, P. Pham, Y. Largman, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1096–1104. Curran Associates, Inc., 2009.

[12] B. T. Meyer and B. Kollmeier. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication*, 53(5):753–767, 2011.

[13] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), ACM MM*, Brisbane, Australia, October 2015.

[14] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of Face & Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Shanghai, China, April 2013.

[15] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013.

[16] S. Souli and Z. Lachiri. Environmental sounds classification based on visual features. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 459–466. Springer, 2011.

[17] L. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg center for Cognition and Communication, Tilburg University, 2009.

[18] T. Wu, Y. Yang, Z. Wu, and D. Li. Masc: a speech corpus in mandarin for emotion analysis and affective

speaker recognition. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1–5. IEEE, 2006.

[19] H. Yin, V. Hohmann, and C. Nadeu. Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency. *Speech Communication*, 53(5):707–715, 2011.