Wavelet-Based Time-Frequency Representations for Automatic Recognition of Emotions from Speech

J. C. Vásquez-Correa[1*], *T. Arias-Vergara*[1], *J. R. Orozco-Arroyave*[1,2], *J. F. Vargas-Bonilla*[1], *E. Nöth*[2]

[1] Faculty of Engineering, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.
[2] Pattern Recognition Lab., Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Email: {jcamilo.vasquez}@udea.edu.co

Abstract

The interest in emotion recognition from speech has increased in the last decade. Emotion recognition can improve the quality of services and the quality of life of people. One of the main problems in emotion recognition from speech is to find suitable features to represent the phenomenon. This paper proposes new features based on the energy content of wavelet based time-frequency (TF) representations to model emotional speech. Three TF representations are considered: (1) the continuous wavelet transform, (2) the bionic wavelet transform, and (3) the synchro–squeezed wavelet transform. The classification is performed using GMM supervectors. Different classification problems are addressed, including high vs. low arousal, positive vs. negative valence, and multiple emotions. The results indicate that the proposed features are useful to classify high vs. low arousal emotions, and that the features derived from the synchro–squeezed wavelet transform are more suitable than the other two approaches to model emotional speech.

Keywords: Speech emotion recognition, Continuous Wavelet Transform, Bionic Wavelet Transform, Synchro – Squeezing, Time–frequency analysis, GMM-supervectors.

1 Introduction

The interest in emotion recognition has been increased in the field of speech and language processing over the last decade [1]. Most of the technological applications related to emotion recognition include support in call centers, tutoring systems, public surveillance, and psychological treatment, among others. One of the main challenges is to find the feature sets that provides the best representation of the emotional speech. For such purpose, new features may be proposed to improve the performance of the systems. Currently, the feature sets used are formed with large sets of acoustic features and include measures derived from prosody, spectral and cepstral features such as Mel frequency cepstral coefficients (MFCC), and voice quality measures [2]. In [3] the authors use the 384 features used in the "INTERSPEECH 2009 emotion challenge" [4] to classify high vs. low arousal emotions, positive vs. negative valence emotions, and multiple emotions in different databases such as Berlin [5] and enterface05 [6]. The authors use a support vector machine (SVM), and follow a leave one speaker out (LOSO) cross–validation. The reported accuracies are around to 96% in Berlin and 76% in enterface05 for the detection of high vs. low arousal, 80% in Berlin and 65% in enterface05 for the positive vs. negative valence, and 80% in Berlin for the classification of seven emotions, and 68% in enterface05 for the classification of six emotions. The same feature set was considered in [7] to classify anger, sadness, happiness, and neutrality in IEMOCAP database [8]. The authors propose a new classification scheme based on a hierarchical binary decision tree using a SVM. The reported unweighted average recall (UAR) considering a speaker independent validation strategy is of up to 58.4%. In [9], the set of 1582 features used as baseline in the "2010 INTERSPEECH computational paralinguistic challenge" was used by the authors to classify the six emotions of the enterface05 database, and four emotions of the FAU-Aibo database, including emphatic, neutral, motherese, and negative emotions. The authors propose a method based on least square regression and report a UAR of 69.3% and 60.5% in enterface05 and in FAU-Aibo, respectively. Recently, feature extraction methods based on the time–frequency (TF) analysis have been successfully applied to classify several emotions and other paralinguistic phenomena from speech [10, 11]. The Wavelet transform has been typically used for TF analysis and to model emotional speech [11–13]. In [12] the authors propose a new set of features based on the energy entropy calculated upon selected bands of the wavelet packet transform (WPT) computed from speech and glottal signals. The authors classify the seven emotions of the Berlin database using a Gaussian mixture model (GMM), and the obtained accuracy is 54%. In [11] the authors use features related to MFCC, linear prediction cepstral coefficients, perceptual linear prediction gamma-tone filter outputs, timbral texture, and energy and entropy measures computed from the WPT. The features are extracted from the speech and glottal signals to model seven emotions of the SAVEE database [14]. The authors propose a feature selection approach based on particle swarm optimization and a classifier based on extreme learning machines, and report an accuracy of 75.4%.

In this paper, features based on the energy content of wavelet based TF representations are proposed for the classification of emotions from speech. We consider three different TF representations: (1) continuous wavelet transform (CWT), (2) bionic wavelet transform (BWT), which is based on a model of the active auditory system [15], and (3) synchro–squeezed wavelet transform (SSWT), which is defined to combine the wavelet analysis and auditory–nerve models [16]. The results obtained with the proposed approach are compared to those obtained with the standard feature set used in the "2009 INTERSPEECH emotion challenge", which is formed with 384 acoustic features. The classification is performed in two stages. The first one consists of a multi-class SVM trained with GMM supervectors formed by concatenating the mean vectors of GMMs trained with the feature sets. The second one consists of taking the distances to the hyper–plane obtained from the first stage and use them as new features to train a second multi–class SVM. Four different datasets for emotion recognition from speech are considered: The Berlin, enterface05, IEMOCAP, and SAVEE databases. Three different classification tasks are also considered: the detection

of (1) high vs. low arousal emotions, (2) positive vs. negative valence emotions, and (3) the recognition of multiple emotions: seven emotions in Berlin, six in enterface05, four in IEMOCAP, and seven in SAVEE. According to our results, the features from the SSWT provide to be the most useful to characterize the emotional speech, rather than the CWT and BWT, particularly in the detection of high vs. low arousal emotions. The rest of the paper is as follows: section 2 contains the description about the methods for feature extraction and classification. Section 3 describes the databases used and the experimental setup. Section 4 contains the description of the results. Finally section 5 includes the main conclusions derived from this study.

2 Material and methods

The methodology followed in this study comprises four stages. (1) The speech utterances are segmented into voiced and unvoiced segments. (2) The three different TF representation are computed for each segment, separately. (3) The energy content in different frequency bands is calculated for each representation forming the feature vectors. (4) Finally the features are used to train the classification scheme based on GMM supervectors.

2.1 Segmentation

Voiced and unvoiced frames are segmented from the speech signals using Praat [17]. This segmentation process has been successfully used in the automatic recognition of emotions and other paralinguistic tasks [18, 19].

2.2 Time-frequency representations

CWT: The CWT is introduced as an alternative to represent and decompose non-stationary signals. The CWT allows a TF multi–resolution analysis based on the decomposition of the signal into time–variable length frames. In the CWT the base functions $\psi_{s,u}(t)$ are small waves of limited duration known as wavelets, whose energy is located around a fixed point. These waves are scaled and translated to create a complete base of the decomposition space. Formally, the CWT of a signal $x(t)$ is given by Equation 1. Where s defines the scale, and u the translation.

$$\mathrm{CWT}_{\mathrm{x}(u,s)} = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt \qquad (1)$$

BWT: The BWT is a TF representation derived from CWT based on a model of the active auditory system [15]. This transform has been widely used to design cochlear implants and speech enhancement algorithms [20]. Formally, the BWT is a time adaptive Wavelet transform designed especially to model speech signals with the Morlet Wavelet function. The BWT is defined by Equation 2 [15].

$$\mathrm{BWT}_{\mathrm{x}(u,s)} = \int_{-\infty}^{\infty} x(t) \frac{1}{\lambda\sqrt{s}} \psi^* \left(\frac{t-u}{s\lambda} \right) \qquad (2)$$

The main difference between the BWT and the CWT is the introduction of the time–adaptive parameter λ. The function λ is derived from the active auditory model and it is defined by Equation 3. Where α is a saturation constant, and β and γ are the gains of the model. In this study, values of $\alpha = 0.8$, $\beta = 0.87$, and $\gamma = 0.45$ are considered, as in related work [15, 20].

$$\lambda = \frac{1}{1 - \alpha \frac{\beta}{\beta + |\mathrm{BWT}_{\mathrm{x}}(u,s)|}} \cdot \frac{1}{1 + \gamma \left| \frac{\partial}{\partial t} \mathrm{BWT}_{\mathrm{x}}(u,s) \right|} \qquad (3)$$

SSWT: The SSWT was introduced to incorporate the Wavelet transform and auditory nerve–models into a TF representation to model speech signals [16]. The aim of the synchro–squeezing is to "sharpen" the CWT by "re–allocating" the value of the point (t, f) in the TF plane into a different point (t', f') according to the local behaviour of the CWT [21]. The aim of SSWT is to obtain a concentrated TF representation of the signal, from which frequency components can be extracted [21]. The $\mathrm{SSWT}_{\mathrm{x}}(u, f)$ is estimated from the $\mathrm{CWT}_{\mathrm{x}}(u, s)$ using Equation 4 [16, 22]. f_i is the frequency index of the $\mathrm{SSWT}_{\mathrm{x}}(u, f)$ and $f_{(s,u)}$ are the instantaneous frequencies where $\mathrm{CWT}_{\mathrm{x}(u,s)} \neq 0$ [22].

$$\mathrm{SSWT}_{\mathrm{x}(u,f_i)} = (\Delta f)^{-1} \sum_{s:\left|f_{(s,u)} - f_i\right| \leq \frac{\Delta f}{2}} \mathrm{CWT}_{\mathrm{x}(u,s)} s^{\frac{-3}{2}} (\Delta s) \qquad (4)$$

Figure 1 illustrates the difference between the three wavelet based TF representations. Note that the frequency components are more spread out in the CWT than in the BWT and in the SSWT.

2.3 Feature extraction

The computation of features consists of dividing the TF representations into 22 frequency regions according to the Bark scale. Each region corresponds to sub-band frequencies from 0 to 8 kHz. The energy content of each band is extracted using Equation 5. u_k and f_i are the time, and frequency index of each representation, respectively. f_i is calculated according to the Bark scale using Equation 6. WT corresponds to any of the described transformations: CWT, BWT, or SSWT.

$$E[i] = log \left| \frac{1}{N} \sum_{f_i} \sum_{u_k}^{N} \left| \mathrm{WT}_{(u_k, f_i)} \right|^2 \right| \qquad (5)$$

$$f_i = 13 \cdot \arctan(0.00076 f) + 3.5 \arctan\left(\left(\frac{f}{7500} \right)^2 \right) \qquad (6)$$

The speech segments are down–sampled to 16 kHz to avoid sampling frequency dependent results. Then, each TF representation is calculated upon frames of 40 ms length and 20 ms time–shift. Figure 2 summarizes the feature extraction process.

2.4 Modeling and Classification

The features extracted from the voiced and unvoiced segments are modeled separately using GMM supervectors, which afterwards are considered for training two SVMs. Finally, the distances to the SVMs hyperplanes are used as features to train a second SVM to make the final decision. The process is detailed in the following subsections. For the muti–class experiments, the SVMs are trained following the one-vs-one strategy.

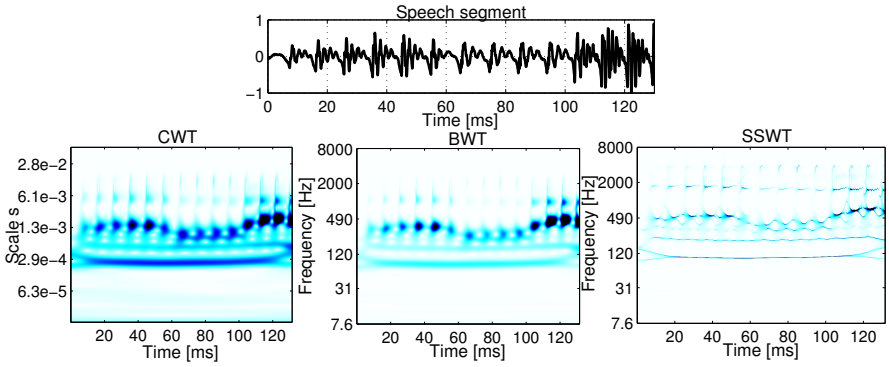

Figure 1: Wavelet based TF representations. Speech signal (up), CWT (left), BWT (middle), SSWT (right)

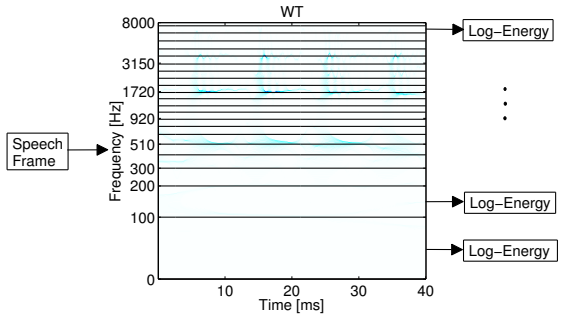

Figure 2: Feature extraction process considering the time-frequency representations

2.4.1 GMM-UBM and supervectors

A GMM is defined as a probabilistic model represented by the linear combination of several multivariate Gaussian components. The GMM is defined using Equation 7, where M is the number of Gaussian components, P_j corresponds to the prior probability of the $j-$th component, and $\mathcal{N}$ is a multivariate Gaussian density function with mean vector μ_j and covariance matrix $\mathbf{\Sigma}_j$.

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{M} P_j \mathcal{N}(\mathbf{x}|\mu_j, \mathbf{\Sigma}_j) \tag{7}$$

Training a GMM consists of estimating the parameters $\Theta = \{P, \mu, \mathbf{\Sigma}\}$ from a training set. The most common method for estimating these parameters is the Expectation Maximization (EM) algorithm [23]. The UBM is a speaker independent model trained with the EM algorithm with a set of samples from all classes, e.g, emotions, and a large number of speakers from the training set [23]. After training the UBM, individual GMMs for each emotion are created following the maximum a posterior (MAP) rule. The GMM supervector is formed by concatenating the mean vectors μ_j of the adapted GMMs [24]. For this study, each supervector is used as a new feature to train a SVM with a Gaussian kernel. This method comprises a "hybrid" classification strategy, where the generative GMM-UBM model is used to create new feature vectors for a discriminative classifier. The number of Gaussian components M is optimized in a grid-search from 2 to 8 in steps of 1. The complexity parameter C and the bandwidth of the kernel γ of the SVM are also optimized through a grid-search in powers of ten with $10^{-1} < C < 10^4$, and $10^{-2} < \gamma < 10^2$.

2.4.2 Second classification stage

The second classification stage consists of fusing the scores (the distances to the separating hyperplanes of the SVMs) obtained from the first stage using the voiced and unvoiced features separately. Those scores are fused and used to train another SVM with a Gaussian kernel to take the decision about which emotion is detected on each utterance.

3 Experimental framework

3.1 Datasets

Berlin emotional database [5]: It contains 534 utterances produced by 10 German native speakers who acted 7 different emotions including anger, disgust, fear, happiness, sadness, boredom, and neutral.

enterface05 database [6]: It contains 1317 audiovisual recordings with 6 emotions produced by 44 speakers, including anger, disgust, fear, happiness, sadness, and surprise. Each subject listened six successive short stories. After each story the subject had to react to the situation by reading predefined sentences closely related to each story.

IEMOCAP database [8]: The interactive emotional dyadic motion capture (IEMOCAP) database contains approximately 12 hours of audiovisual data, including video, speech, motion capture of the face, and text transcriptions. The audio files consist of 10039 utterances produced by 10 English native speakers who acted 10 different emotions. The database consists of dyadic sessions where actors performed improvisations or scripted scenarios, specifically selected to elicit emotional expressions.

SAVEE database [14]: The surrey audiovisual expressed emotion (SAVEE) database consists of utterances from 4 male actors in 7 different emotions such as anger, disgust, fear, happiness, sadness, surprise, and neutral. The database is formed by 480 British English utterances.

3.2 Experimental setup

Three experiments are addressed using the recordings from each database. We perform the classification of (1) high vs. low arousal emotions, (2) positive vs. negative valence emotions, and (3) all emotions from the databases: seven emotions in Berlin, six in enterface05, four in IEMOCAP, and seven in SAVEE. Table 1 lists the emotions considered for the three experiments addressed in this study.

Table 1: Experiments addressed in this study

Database	2-class Arousal	2-class Valence	multi-class All emotions
Berlin	*High*: Fear, Disgust, Happiness, Anger. *Low*: Boredom, Neutral, Sadness.	*Positive*: Neutral, Happiness. *Negative*: Boredom, Anger Sadness, Fear, Disgust.	Fear, Disgust Happiness, Neutral Boredom, Sadness Anger
Enterface05	*High*: Fear, Disgust, Happiness, Anger. Surprise *Low*: Sadness.	*Positive*: Surprise Happiness. *Negative*: Anger Sadness, Fear, Disgust.	Fear, Disgust Happiness, Anger Surprise, Sadness
IEMOCAP	*High*: Fear, Disgust, Happiness, Anger. Surprise, Excitation, Frustration *Low*: Sadness, Neutral.	*Positive*: Surprise, Neutral Happiness, Excitation *Negative*: Anger, Frustration, Sadness, Fear, Disgust.	Anger, Sadness Happiness, Anger
SAVEE	*High*: Fear, Disgust, Happiness, Anger. Surprise *Low*: Neutral, Sadness.	*Positive*: Surprise, Neutral Happiness. *Negative*: Anger Sadness, Fear, Disgust.	Fear, Disgust Happiness, Anger Surprise, Sadness Neutral

3.3 Validation

A speaker independent cross–validation strategy based on LOSO is followed. The performance measure used for the evaluation of the methodology is the UAR due to the high unbalanced data. UAR is defined as the unweighted average of the class–specific recalls achieved by the system.

4 Results and Discussion

The results with the proposed features are compared to those obtained with the feature set used as baseline in the "INTERSPEECH 2009 emotion challenge" [4], which is formed with 384 acoustic features computed with the OpenEAR toolkit. Table 2 displays the results for the detection of low vs. high arousal. The results obtained with the fusion scheme are better than those obtained with each kind of segment separately, except for the results in Berlin, where the highest UAR is obtained with the voiced features. Note also that in most of the cases the wavelet–based TF representations provide higher UARs than OpenEAR.

Table 2: Detection of high vs. low arousal emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	95.7±5.6	82.5±9.1	81.2±2.2	74.4±3.8
	U	89.1±8.7	79.8±8.1	79.6±1.2	75.1±2.7
	Fusion	93.3±8.3	87.3±7.4	80.8±2.5	76.4±2.5
BWT	V	95.6±5.5	82.3±8.0	81.5±1.7	74.3±4.2
	U	89.6±8.5	80.4±7.2	79.8±1.5	74.8±2.8
	Fusion	94.0±6.6	84.6±7.1	81.9±2.2	76.1±4.0
SSWT	V	95.8±5.5	84.4±8.3	81.1±1.7	75.7±4.7
	U	89.2±8.4	79.5±6.7	80.4±1.4	75.6±2.9
	Fusion	95.0±5.5	81.8±5.7	80.2±2.9	77.2±3.6
OpenEAR	-	97.3±3.0	83.3±8.8	81.0±2.0	75.5±3.8

Table 3 contains the results classifying positive vs. negative valence emotions. In general, the highest UARs are obtained with OpenEAR. This is likely due to the fact that we consider only features based on the energy content of the TF representation, which may not provide enough information to model the valence dimension. The SSWT provides better results than the BWT and CWT in three of the four databases. On the other hand, the fusion scheme in SAVEE and IEMOCAP improves the results when voiced and unvoiced features are used separately. Table 4 contains the results for the classification of multiple emotions. In general, the highest results are obtained also with OpenEAR. Note also that in all of the cases the fusion scheme improves the results relative to those obtained when voiced and unvoiced segments are modeled separately.

Table 3: Detection of positive vs. negative valence emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface05	IEMOCAP
CWT	V	80.0±3.7	64.4±5.0	74.6±1.7	54.5±3.8
	U	76.3±5.4	63.8±3.2	73.4±2.6	57.5±2.3
	Fusion	78.2±4.2	66.7±3.5	74.4±2.0	58.4±4.7
BWT	V	80.0±3.7	63.8±6.3	74.2±2.0	54.6±3.6
	U	76.4±6.7	63.8±4.5	73.6±2.7	57.6±2.1
	Fusion	78.0±5.5	64.6±5.9	73.5±4.2	58.1±3.2
SSWT	V	81.7±4.6	64.2±4.8	75.6±2.9	56.2±4.0
	U	76.9±6.0	63.1±3.4	74.3±2.8	58.3±1.9
	Fusion	78.5±3.8	65.4±5.3	73.8±3.6	59.5±3.3
OpenEAR	-	87.2±2.4	72.5±5.7	81.4±3.6	59.0±3.2

Table 4: Classification of multiple emotions. V: voiced, U: unvoiced.

Features	Segm.	Berlin	SAVEE	enterface-05	IEMOCAP
CWT	V	61.3±8.3	40.6±13.5	48.4±4.7	46.7±6.0
	U	54.7±6.6	39.4±5.8	45.7±4.0	51.3±3.6
	Fusion	66.6±6.5	43.8±9.0	51.3±5.6	55.9±5.0
BWT	V	63.7±9.1	41.2±14.9	48.4±4.4	46.6±5.3
	U	55.5±7.4	39.8±4.3	44.9±4.3	51.2±3.9
	Fusion	66.5±6.5	47.3±10.3	49.7±4.3	55.2±5.7
SSWT	V	64.0±8.0	42.7±11.1	48.0±3.5	48.7±5.0
	U	55.0±8.2	39.6±6.2	45.9±3.6	52.0±2.9
	Fusion	69.3±7.6	45.4±12.1	48.8±5.8	58.2±4.1
OpenEAR	-	80.4±8.0	49.4±17.6	63.2±6.7	57.2±2.8

5 Conclusions

This study evaluates three different wavelet based TF representations to model emotional speech. Three classification problems are addressed: detection of high vs. low arousal emotions, classification of positive vs. negative valence emotions, and the recognition of multiple emotions. The emotions are modeled considering a scheme based on GMM supervectors that are used to train a discriminative classifier based on a SVM. The TF representations include different versions of the wavelet transform: the CWT, the BWT, and the SSWT. When comparing these three TF–based transformations, SSWT provides better results, indicating that the re–allocating method that sharpens the frequency components of the spectrum to a narrower band seems to be useful to model emotional speech. The proposed features are computed separately for voiced and unvoiced segments. In most of the cases the highest UARs are obtained with the features extracted from voiced segments. The fusion scheme shows to be useful to combine the information provided by both kinds of segments. The results with the proposed approach are better than those obtained with openEAR when classifying high vs. low arousal emotions. This could be explained due to the fact that we only extract features based on the energy content of the TF representation, which is useful for the detection of arousal, but not to classify valence or multiple emotions. Further experiments shall be performed considering other descriptors extracted from the TF representations to improve the results.

Acknowledgment

This work is partially funded by COLCIENCIAS through the project # 111556933858.

References

[1] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion recognition in naturalistic speech and language - a survey," in *Emotion Recognition: A Pattern Analysis Approach*, pp. 237–267, Wiley, 2015.

[2] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[3] F. Eyben, A. Batliner, and B. Schuller, "Towards a standard set of acoustic features for the processing of emotion in speech," in *Proceedings of Meetings on Acoustics*, vol. 9, pp. 1–12, 2010.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the International conference of the speech and communication associatoin INTERSPEECH*, pp. 312–315, 2009.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *Proceedings of the International conference of the speech and communication associatoin INTERSPEECH*, pp. 1517–1520, 2005.

[6] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proceedings of International Conference on Data Engineering Workshops*, pp. 8–15, 2006.

[7] C. C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[8] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[9] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[10] T. Villa-Cañas, J. D. Arias-Londoño, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, and E. Nöth, "Low-frequency components analysis in running speech for the automatic detection of Parkinson's disease," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[11] H. Muthusamy, K. Polat, and S. Yaacob, "Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals," *PloS one*, vol. 10, no. 3, p. e0120344, 2015.

[12] L. He, M. Lech, J. Zhang, X. Ren, and L. Deng, "Study of wavelet packet energy entropy for emotion classification in speech and glottal signals," in *Fifth International Conference on Digital Image Processing*, 2013.

[13] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Non-linear dynamics characterization from wavelet packet transform for automatic recognition of emotional speech," in *Recent Advances in Nonlinear Speech Processing*, pp. 199–207, Springer, 2016.

[14] S. Haq and P. J. B. Jackson, "Multimodal emotion recognition," *Machine audition: principles, algorithms and systems, IGI Global, Hershey*, pp. 398–423, 2010.

[15] J. Yao and Y. T. Zhang, "Bionic wavelet transform: a new time-frequency method based on an auditory model," *IEEE Transactions on Biomedical Engineering* , vol. 48, no. 8, pp. 856–863, 2001.

[16] I. Daubechies and S. Maes, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets in medicine and biology*, pp. 527–546, 1996.

[17] P. Boersma and D. Weenik, "PRAAT: a system for doing phonetics by computer. Report of the Institute of Phonetic Sciences of the University of Amsterdam.," 1996.

[18] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[19] J. C. Vásquez-Correa, N. Garcia, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Emotion recognition from speech under environmental noise conditions using wavelet decomposition," in *Proceedings of the International Carnahan Conference on Security Technology (ICCST)*, pp. 1–5, 2015.

[20] S. M. Govindan, P. Duraisamy, and X. Yuan, "Adaptive wavelet shrinkage for noise robust speaker recognition," *Digital Signal Processing*, vol. 33, pp. 180–190, 2014.

[21] I. Daubechies, J. Lu, and H. T. Wu, "Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 243–261, 2011.

[22] G. Thakur, E. Brevdo, N. S. Fučkar, and H. T. Wu, "The synchrosqueezing algorithm for time-varying spectral analysis: robustness properties and new paleoclimate applications," *Signal Processing*, vol. 93, no. 5, pp. 1079–1094, 2013.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[24] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.