Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification

L. Zão, *Student Member, IEEE*, D. Cavalcante, and R. Coelho, *Member, IEEE*

***Abstract*—In this letter, the pH time-frequency vocal source feature is proposed for multistyle emotion identification. A binary acoustic mask is also used to improve the emotion classification accuracy. Emotional and stress conditions from the Berlin Database of Emotional Speech (EMO-DB) and Speech under Simulated and Actual Stress (SUSAS) databases are investigated in the experiments. In terms of emotion identification rates, the pH outperforms the mel-frequency cepstral coefficients (MFCC) and a Teager-Energy-Operator (TEO) based feature. Moreover, the acoustic mask achieves accuracy improvement for both the MFCC and the pH feature.**

***Index Terms*—Binary acoustic mask, Hurst exponent, pH feature, speech emotion recognition.**

I. Introduction

THE effect of emotional expression on speech is an interesting issue and it has been the object of many studies [1]–[4]. The emotional state affects the speech production by introducing changes in muscle tension and in the breathing rate. The identification of emotions from speech is less intrusive than other approaches, such as the heartbeat rate and blood pressure measures. In [1], it was found that high-activation emotions, such as anger and happiness, induce an arousal of the sympathetic nervous system. The energy of the resulting speech signal is more concentrated at the high frequency components. On the other hand, low-arousal emotions, like boredom and sadness, produce low-pitched speech signals. Thus, the emotion acoustic evidence is more likely to be found in the voiced segments of speech. The absence of emotion leads to a neutral speech.

The search for speech features suitable for emotion classification is still a crucial task. Vocal source features, extracted from residual speech signals, have valuable information about the pitch harmonics distribution [5]. The pitch carries important information about emotion since it depends on the vocal folds tension. Moreover, the excitation source contribution on the short-time speech spectral envelope is affected by the speaking style [6]. For instance, forceful and abrupt speech with higher frequency energy presents power spectral density (PSD) with a -9 dB/octave roll-off. On the other hand, more relaxed speech with dominance of low frequency energy presents -15 dB/octave. In neutral speech it is observed an average decaying rate of -12 dB/octave. Due to their discriminating capability, vocal source features have been investigated for automatic speech emotion recognition [4], [7].

In this letter, the pH time-frequency vocal source feature [8] is proposed for automatic multistyle speech emotion classification. The pH consists of a vector of Hurst exponent (H) values and it is closely related to the excitation source. Another contribution of this letter is the introduction of a binary acoustic mask to improve the multistyle emotion recognition. The amplitude modulation spectrogram (AMS) and the Gaussian mixture models (GMM) are adopted for the masking procedure. In the literature, the use of binary masks are mainly focused on the speech classification [9] and the speech intelligibility improvement [10] in background noisy scenarios. In this work, it is used to identify or select the speech spectro-temporal components that are most related to the emotional states.

The speech emotion identification experiments are conducted with a set of emotions from the Berlin Database of Emotional Speech (EMO-DB) [11] and stress conditions from the Speech under Simulated and Actual Stress (SUSAS) [12] databases. The accuracy of the pH vector as vocal source emotion feature is firstly compared to those obtained with the MFCC and the Critical Band TEO Autocorrelation Envelope (TEO-CB-Auto-Env) [2]. The results show that the pH outperforms the baseline features for both speech databases. Finally, the proposed acoustic mask is also applied to improve the emotion classification accuracy. The best recognition results are obtained with the pH feature applied to the masked signals.

II. pH Vocal Source Feature

The pH is a time-frequency feature that was proposed for speaker identification and verification systems [8]. The Hurst exponent ($0 < H < 1$) expresses the time-dependence or scaling degree of the speech signal $y(t)$, whose autocorrelation coefficient function (ACF) asymptotically decays according to

$$\rho(k) \sim H(2H-1)k^{2(H-2)}, \quad k \to \infty. \tag{1}$$

The H values can be related to the spectral characteristics of $y(t)$. The pH is here proposed to represent the speech emotional states [6], as follows:

- High-arousal emotions ($0 < H < 1/2$): the dominant high frequency components (-9 dB/octave roll-off) induce the ACF to rapidly decay to zero.

Manuscript received November 21, 2013; accepted February 26, 2014. Date of publication March 12, 2014; date of current version March 21, 2014. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under the Grant 304254/2012-6. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Paris Smaragdis.

L. Zão and D. Cavalcante are with the Graduate Program in Defense Engineering, Military Institute of Engineering (IME), Rio de Janeiro, Brazil (e-mail: zao@ime.eb.br, dirceu_cavalcante@ime.eb.br).

R. Coelho is with the Laboratory of Acoustic Signal Processing, Electrical Engineering Department, Military Institute of Engineering (IME), Rio de Janeiro, Brazil (e-mail: coelho@ime.eb.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2311435

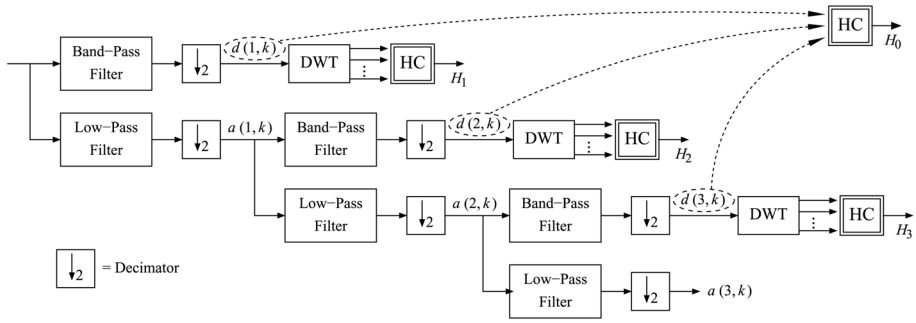

Fig. 1. An example of a pH vector estimation using the M-dim-wavelets with 3 decomposition stages.

- Neutral speech ($H \approx 1/2$): the ACF usually exhibits exponential decay and the PSD decaying rate is about -12 dB/octave.
- Low-arousal emotions ($1/2 < H < 1$): the low-frequency energy leads to a slowly vanishing ACF, with -15 dB/octave PSD roll-off.

A. pH Vector Extraction

The wavelet-based multi-dimensional estimator (M-dim-wavelets) [8] was proposed as the pH feature extractor and is based on the method described in [13]. The estimation procedure is as follows:

- Wavelet decomposition: apply the discrete wavelet transform (DWT) to successively decompose a sequence of samples into approximation ($a(j,k)$) and detail ($d(j,k)$) coefficients, where j is the decomposition scale ($j = 1, 2, \ldots, J$) and k is the coefficient index of each scale.
- Hurst exponent computation (HC) [13]: for each scale j, the variance $\sigma_j^2 = (1/n_j)\sum_k d(j,k)^2$ is evaluated from the detail coefficients, where n_j is the number of available coefficients for each scale j. In [13], it is shown that $E[\sigma_j^2] = \mathcal{C}_H j^{2H-1}$, where $\mathcal{C}_H$ is a constant. A weighted linear regression is then used to obtain the slope α of the plot of $y_j = \log_2(\sigma_j^2)$ versus j. The value of H is given by $H = (1+\alpha)/2$.
- pH vector composition: the pH vector is composed of $(J+1)$ values of $H[H_0, H_1, \ldots, H_J]$. The H_0 component is computed from the decomposition of the entire speech signal. The other values $(H_1, H_2, \ldots, H_J)$ are obtained after re-applying the DWT decomposition to each of the J detail sequences. Fig. 1 shows an example of the pH estimation considering $J = 3$ decomposition stages, i.e., $[H_0, H_1, H_2, H_3]$

During the pH extraction, the time-frequency multi-resolution analysis captures the higher order correlations of the speech samples. Such correlations are also present in the vocal sources features, since they are extracted from the linear prediction residual. Thus, the pH feature is closely related to the excitation source, which is very useful for emotion classification [2].

Fig. 2 illustrates the H values distribution of the speech signals collected from EMO-DB corresponding to four different emotions: anger, happiness, neutral and sadness. The Hurst exponent is computed with the wavelet-based estimator [13] from non-overlapping speech segments of 32 ms. In this work, the

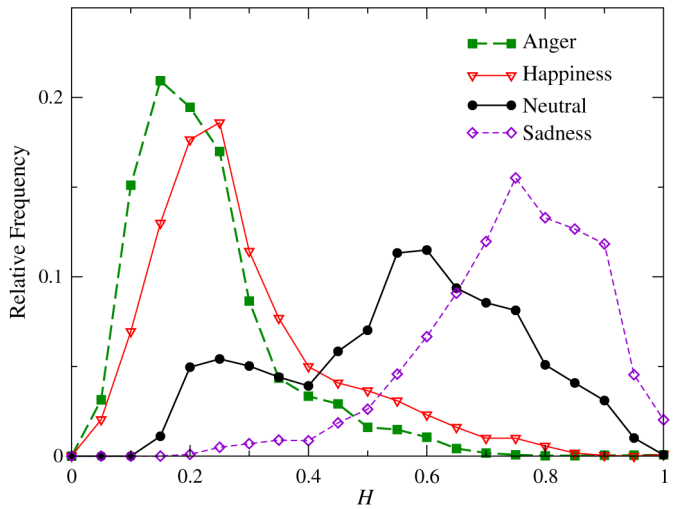

Fig. 2. Distribution of the Hurst values for speech samples under four emotional states.

Daubechies wavelets filters are applied for the DWT decomposition. It can be seen that the H values have higher relative frequencies in the range $0 < H < 1/2$ for the high-arousal emotions, i.e., anger and happiness. On the other hand, H values for sadness are mostly concentrated around values of $H \geq 1/2$. This emphasizes that the pH feature is able to discriminate the acoustic emotions.

III. AMS-GMM Acoustic Mask

The proposed acoustic mask uses GMM to model the amplitude modulation spectrograms and decide which components should be removed from the speech signal. The AMS was proposed in [14] to compute the envelope spectrum of bandpass-filtered speech signals and capture the modulation frequency variations of each frequency band. Since the modulation frequency pattern observed in voiced speech waveforms reflects the speaker's emotional state [2], the AMS is able to detect the speech signal components that are most related to the target emotional state. Fig. 3 shows the block diagram of the multistyle emotion classification using the AMS-GMM acoustic mask and four example emotional states.

The AMS estimation begins with the decomposition of the speech signals into 25 sub-bands according to the mel-frequency scale. The envelopes in each sub-band are computed and divided into short-time segments of 32 ms using 50%-overlapping Hanning windows. The fast Fourier transform (FFT) is then applied to estimate the modulation frequency spectrum

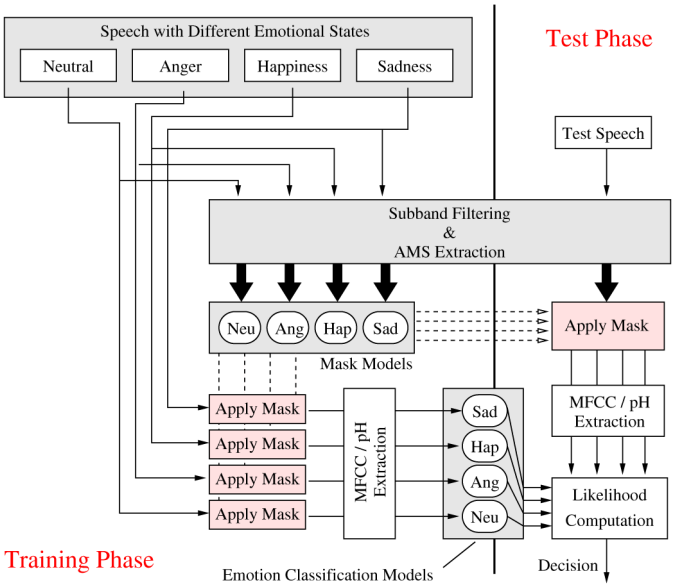

Fig. 3. The emotion classification schematic using the acoustic mask considering four example emotional states. The emotional states are replaced by the stress conditions when adopted for the SUSAS database.

of each sub-band. Finally, each spectrum is divided into 15 channels uniformly distributed in the range 15.6 Hz–400 Hz [10]. The AMS vector $\vec{x}_{b,q}$ of sub-band b and time frame q is composed with the FFT magnitudes from each of the 15 channels. The AMS matrix is obtained by $\mathbf{X}_b = [\vec{x}_{b,1}\vec{x}_{b,2}\cdots\vec{x}_{b,Q}]$, where Q is the number of frames.

A. Training Phase

For each emotional state $\mathcal{E}$, the AMS matrices estimated from the training utterances are used to obtain 25 Gaussian mixture models $\lambda_b^{\mathcal{E}}$, one for each sub-band. The next step is to use these mask models to decide whether to eliminate the spectro-temporal components of the training speech signals. Thus, for each (non-neutral) emotional state $\mathcal{E}$, the binary mask $M_{\mathcal{E}}(b,q)$ of sub-band b and frame q is estimated as

$$M_{\mathcal{E}}(b,q) = \begin{cases} 1, & \text{if } p(\vec{x}_{b,q}|\lambda_b^{\mathcal{E}})/p(\vec{x}_{b,q}|\lambda_b^{N}) > \theta_b; \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In (2), $p(\vec{x}_{b,q}|\lambda_b^{\mathcal{E}})$ and $p(\vec{x}_{b,q}|\lambda_b^{N})$ are the likelihood functions calculated as sums of Gaussian densities, the superscript index N refers to the neutral emotional state and θ_b is the mask threshold for sub-band b. It means that, in the speech reconstruction, the mask removes the spectro-temporal regions (i.e., $M_{\mathcal{E}}(b,q) = 0$) which are more likely to belong to neutral (compensated by a factor θ_b) than to emotion $\mathcal{E}$. For neutral speech, the decision criteria is given by

$$M_N(b,q) = \begin{cases} 1, & \text{if } p(\vec{x}_{b,q}|\lambda_b^{N})/p(\vec{x}_{b,q}|\lambda_b^{\bar{N}}) > \theta_b; \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $\lambda_b^{\bar{N}}$ is the mask model of sub-band b obtained from the AMS matrices of all emotional states, except neutral.

After obtaining the mask GMMs, the acoustic mask is applied to reconstruct the training utterances retaining the spectro-temporal regions most related to their corresponding emotional states. Then, speech feature matrices are extracted from the masked signals and the GMM $\Lambda_{\mathcal{E}}$ is finally obtained for each emotional state $\mathcal{E}$.

B. Test Phase

During tests, the AMS matrices of the input speech signal are estimated after the sub-band decomposition. Then, multiple masked versions of the speech signal are reconstructed by applying the criteria defined in (2) and (3) for each emotional state. Let $\mathbf{Y}_{\mathcal{E}}$ represent the speech feature matrix (MFCC or pH) obtained from the reconstructed signal considering the mask of the emotional state $\mathcal{E}$. Then, the identified emotion $\hat{\mathcal{E}}$ for the input signal is the one that maximizes the likelihood function $p(\mathbf{Y}_{\mathcal{E}}|\Lambda_{\mathcal{E}})$, i.e., $\hat{\mathcal{E}} = \arg\max_{\mathcal{E}} p(\mathbf{Y}_{\mathcal{E}}|\Lambda_{\mathcal{E}})$.

IV. Experimental Setup and Results

The proposed pH feature and GMM-AMS mask are evaluated in acoustic multistyle emotion identification experiments. The leave-one-speaker-out methodology (LOSO) [3] is adopted to achieve speaker independence. The pH vectors are extracted with $J = 5$ decomposition stages considering two sizes of speech segments: 20 ms and 50 ms. Thus, pH vectors are obtained every 10 ms with $2(J+1) = 12$ components. The H values are computed using the Daubechies wavelet filters with 12 coefficients. For the performance comparison, 12-dimensional MFCC vectors are obtained following the same configuration adopted in [3], i.e., with speech frames of 25 ms at a frame rate of 10 ms. The TEO-CB-Auto-Env vectors are obtained from speech frames of 75 ms and 50% overlapping and are composed of 16 coefficients [2]. In order to evaluate the excitation source discriminating power, only the high energy voiced segments are considered in the experiments. Thus, neither Δ nor $\Delta\Delta$ are appended to the feature vectors.

Regarding the tests with the acoustic mask, the threshold values (θ_b, $b = 1, 2, \ldots, 25$) are set as to retain 80% of the regions most related to the corresponding emotional state. It means that in each sub-band, 20% of the spectro-temporal regions are suppressed by the masking procedure. Since it led to the beste results in preliminary experiments, the GMM used for the mask and emotional models are composed of 32 Gaussian densities with diagonal covariance matrices.

A. Results with EMO-DB

The EMO-DB corpus is composed of 494 utterances with archetypical emotions obtained from ten professional actors. Each utterance was obtained with a sampling rate of 16 kHz and was previously approved by a perceptual emotion recognition test. The EMO-DB contains seven emotional states: anger, boredom, disgust, fear, happiness, neutral and sadness.

Tab. I presents the emotion recognition accuracies achieved with pH, MFCC and TEO-CB-Auto-Env. Note that, except for happiness and sadness, the pH feature leads to the highest correct identification rates for all the emotional states. For fear, it is improved from 33% with MFCC to 62% with the pH. In average, the accuracy obtained with the pH feature is 68.1%. This result is 6.8 percentage points (p.p.) higher than that achieved with the MFCC vectors, i.e., 61.3%. Moreover, the TEO-CB-Auto-Env obtains the lowest average result: 50.4%.

The correct identification rates illustrated in Fig. 4 are obtained with the proposed AMS-GMM acoustic mask. The av-

TABLE I
EMOTION RECOGNITION ACCURACIES (%) FOR EMO-DB.

	Actual Emotion	Classified Emotion						
		Ang.	Bor.	Dis.	Fear	Hap.	Neu.	Sad.
pH feature	Anger	**86**	0	2	2	10	0	0
	Boredom	0	**61**	13	2	0	20	4
	Disgust	0	7	**67**	6	10	10	0
	Fear	0	3	5	**62**	16	11	3
	Happiness	25	2	6	8	**48**	11	0
	Neutral	0	17	2	8	0	**71**	2
	Sadness	0	12	0	0	0	6	**82**
	Average classification accuracy: **68.1%**							
MFCC feature	Anger	**85**	0	1	0	14	0	0
	Boredom	0	**53**	8	5	5	25	4
	Disgust	5	0	**61**	5	11	18	0
	Fear	7	13	11	**33**	22	15	0
	Happiness	19	2	8	11	**58**	3	0
	Neutral	0	23	5	5	0	**65**	1
	Sadness	0	6	11	0	0	9	**74**
	Average classification accuracy: **61.3%**							
TEO-CB-Auto-Env	Anger	**60**	3	4	1	27	5	0
	Boredom	5	**58**	9	1	4	19	4
	Disgust	18	13	**40**	0	13	13	3
	Fear	15	6	16	**27**	11	18	7
	Happiness	37	5	8	0	**42**	8	0
	Neutral	6	44	10	1	3	**36**	0
	Sadness	0	2	4	0	0	4	**90**
	Average classification accuracy: **50.4%**							

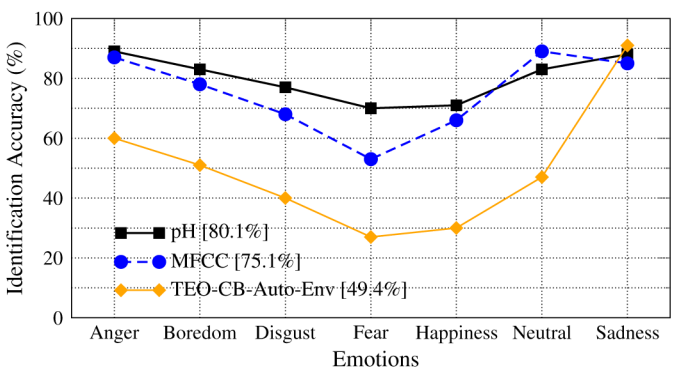

Fig. 4. Emotion identification accuracies (%) for EMO-DB using the AMS-GMM acoustic mask.

erage results are presented in the legends. The adoption of the AMS-GMM mask improves the identification results for the pH and MFCC features. For instance, considering the pH feature, more than 20 p.p. gain is obtained due to the masking procedure for boredom and happiness emotions. The results with TEO-CB-Auto-Env feature are the only ones which are not improved with the acoustic mask. It may be observed that the pH vocal source feature outperforms the MFCC and TEO-CB-Auto-Env for most of the emotions, and also in terms of average recognition accuracy.

B. Results with SUSAS

The SUSAS database is composed of 3593 utterances spoken on a park roller-coaster ride by seven speakers. The spoken text corresponds to 35 English short commands, such as "*no*" and "*brake*". The utterances were recorded with a sampling rate of 8 kHz and are divided into four real stress conditions: neutral, medium stress, high stress and screaming.

The multistyle stress recognition accuracies obtained with the MFCC, pH and TEO-CB-Auto-Env are presented in Tab. II.

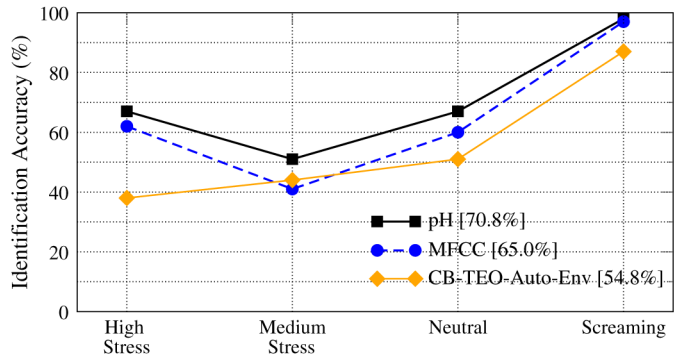

Fig. 5. Stress identification accuracies (%) for SUSAS using the AMS-GMM acoustic mask.

TABLE II
STRESS RECOGNITION ACCURACIES (%) FOR SUSAS.

	Actual Stress Condition	Classified Stress Condition			
		Neutral	Medium	High	Screaming
pH feature	Neutral	**59**	20	20	1
	Medium	29	**36**	35	0
	High	16	22	**62**	0
	Screaming	1	0	0	**99**
	Average classification accuracy: **64.0%**				
MFCC feature	Neutral	**58**	19	23	0
	Medium	25	**36**	39	0
	High	14	33	**53**	0
	Screaming	3	0	0	**97**
	Average classification accuracy: **61.0%**				
TEO-CB-Auto-Env	Neutral	**46**	30	16	8
	Medium	30	**35**	31	4
	High	20	28	**47**	5
	Screaming	5	3	3	**89**
	Average classification accuracy: **54.3%**				

Once again, the pH feature outperforms the baseline features. For the high stress condition, for example, the pH leads to an identification rate of 62%, while 53% is achieved with the MFCC and 47% with TEO-CB-Auto-Env. Fig. 5 shows the stress classification results considering the AMS-GMM mask. Once again, the best performance is achieved with the pH feature extracted from the masked signals. The average identification result with the pH is more than 5 p.p. higher than the MFCC. For both MFCC and pH features, the masking procedure improves the recognition results for three the stress conditions. Considering the use of both proposals (mask and the pH feature), the average identification rate is improved from 61.0% with MFCC to 70.8%.

V. CONCLUSION

This letter proposed the pH time-frequency vocal source feature and a binary acoustic mask for the speech emotion classification. The results show that the pH outperforms the baseline features for the multistyle emotion and stress classification. When compared to the MFCC, the average recognition results obtained with the pH feature achieved absolute improvement of 6.8 p.p. and 3.0 p.p. for the EMO-DB and the SUSAS databases, respectively. Also, the acoustic mask improved the identification rates for the MFCC and pH and both databases. The combination of the pH and the AMS-GMM mask led to more than 18 p.p. gain for the EMO-DB.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, and W. Fellenz, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, pp. 32–80, Jan. 2001.

[2] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 201–216, Mar. 2001.

[3] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *IEEE Workshop on Automatic Speech Recognition Understanding*, 2009, pp. 552–557.

[4] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech recognition: Resources, features and methods," *Patt. Recognit.*, vol. 44, pp. 572–587, Mar. 2011.

[5] N. Wing, P. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE Trans. Speech Audio Process.*, vol. 19, pp. 196–205, Jan. 2011.

[6] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[7] 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 Commun.*, vol. 53, pp. 1062–1087, Nov. 2011.

[8] R. Sant'Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, pp. 931–940, May 2006.

[9] W. Kim and J. Hansen, "A novel mask estimation method employing posterior-based representative mean estimate for missing-feature speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1434–1443, 2011.

[10] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, pp. 1486–1494, Sep. 2009.

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

[12] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. EUROSPEECH'97*, Sep. 1997, vol. 4, pp. 1743–1745.

[13] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Trans. Inf. Theory*, vol. 45, pp. 878–897, Apr. 1999.

[14] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1593–1602, Mar. 1994.