ANALYSIS OF EMOTIONAL SPEECH USING AN ADAPTIVE SINUSOIDAL MODEL

George P. Kafentzis[2], Theodora Yakoumaki[1,2], Athanasios Mouchtaris[1,2], Yannis Stylianou[2,3]

[1]Institute of Computer Science, Foundation for Research and Technology Hellas, Greece
[2]Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
[3]Toshiba Cambridge Research Lab, U.K.

{kafentz, yakumaki}@csd.uoc.gr, mouchtar@ics.forth.gr, yannis.stylianou@crl.toshiba.co.uk

ABSTRACT

Processing of emotional (or expressive) speech has gained attention over recent years in the speech community due to its numerous applications. In this paper, an adaptive sinusoidal model (aSM), dubbed *extended adaptive Quasi-Harmonic Model - eaQHM*, is employed to analyze emotional speech in accurate, robust, continuous, time-varying parameters (amplitude, frequency, and phase). It is shown that these parameters can adequately and accurately represent emotional speech content. Using a well known database of narrowband expressive speech (SUSAS) we show that very high Signal-to-Reconstruction-Error Ratio (SRER) values can be obtained, compared to the standard sinusoidal model (SM). Formal listening tests on a smaller wideband speech database show that the eaQHM outperforms SM from a perceptual resynthesis quality point of view. Finally, preliminary emotion classification tests show that the parameters obtained from the adaptive model lead to a higher classification score, compared to the standard SM parameters.

Index Terms— Extended adaptive quasi-harmonic model, Speech analysis, Emotional speech, Sinusoidal modelling, Emotion classification

1. INTRODUCTION

Emotional (or stressed) speech can be defined as the speech style produced by an emotionally charged speaker. Such speech styles can be characterized as *happy, sad, angry, neutral* and *fearful* speech, among others. Analysis of emotional speech could provide information about the emotional state of the speaker, which can be useful in applications such as health care and emergency conditions, and is a necessary pre-processing step in applications such as recognition and classification. Also, speaker recognition and verification systems could benefit from such an analysis, as well as speech synthesis applications, like unit selection based text-to-speech synthesis or HMM-based speech synthesis.

Numerous approaches have been suggested in the literature in order to show the variation of speech characteristics among different emotion conditions. These variations can form *features* that are exploited to identify and/or classify different emotional speech styles [1]. Womack and Hansen discussed the use of Linear Prediction (LP) coefficients and cepstral features in analyzing and classifying stressed speech [2–5]. Zhou et al [6] have shown that the Teager operator can be used to obtain better results compared to LP-based features in classification of stressed speech. Moreover, it has been suggested that features related to the pitch mean and variance, as well as intensity features, are useful for discrimination among speaking styles [7, 8]. Cummings et al [9] have shown that the glottal pulse shape varies with different stressed conditions. Ruiz et al [10] discussed time and frequency related variabilities in stressed speech, whereas Castellanos et al [11] provided an analysis of general acoustic-phonetic features in Lombard speech. Scherer [12] investigated the intensity, duration, and spectral envelopes in stressed speech for speech and speaker recognition, whereas Bosch [13] has discussed the importance of prosody for emotion recognition in speech. Ramamohan and Dandapat [14] suggested the use of a sinusoidal model (SM) to distinguish between different speaking styles, using its parameters (amplitude, frequency, phase) as features.

In spite of its wide range of applications [15], the Sinusoidal Model (SM) [16] has not been thoroughly engaged in analysis and/or classification of stressed speech until recently [14, 17]. In these approaches, the parameters of sinusoids (amplitude, frequency, and phase) over time are suggested as features for classification or conversion of speech using Hidden Markov Models, Vector Quantization, and Gaussian Mixture Model-based techniques. Although the use of amplitude and frequency contours was straightforward, the phase contours are either disregarded or could not be directly used in the analysis. Furthermore, the parameters obtained from sinusoidal analysis have a significant constraint; they are extracted under the assumption of *local stationarity*, that is, the speech signal is considered as *stationary* inside the analysis window. However, this is not the case for speech styles characterized as *"fast"* or *"angry"*. Recently, the adaptive Sinusoidal Models (aSMs) [18–20] have managed to cope with this problem by projecting the signal onto a set of amplitude- and frequency-varying basis functions *inside* the analysis window. This way, the parameters represent the underlying signal more closely as an AM-FM decomposition. In brief, the adaptive Quasi-Harmonic Model (aQHM) [21] adapts the phase of the basis function to the local characteristics of the signal, whereas the extended adaptive Quasi-Harmonic Model (eaQHM) [19] performs both amplitude and phase adaptation. More recently, the adaptive Harmonic Model (aHM) [20] assumes full-band harmonicity and iteratively adapts the fundamental frequency f_0 to localize harmonics up to the Nyquist frequency. All models have demonstrated their ability to model adequately and accurately speech signals from different languages and different speakers. However, they have not been tested in emotional speech, where it is assumed that the AM-FM components of the speech signal behave differently compared to neutral or conversational speech.

In this paper, the extended adaptive Quasi-Harmonic Model (eaQHM) is utilized to demonstrate its ability to analyze, resynthesize, and classify emotional speech. The speech corpus for the analysis and resynthesis is a high-quality, wideband database containing emotional running speech. Subjective listening tests have been conducted to prove the transparency of the resynthesized speech. It is also shown that eaQHM can efficiently model all styles of emotional speech in this database with high precision, and this is demonstrated

via Signal-to-Reconstruction-Error Ratio (SRER) values, compared to the standard SM. Moreover, an emotion classification task is presented using the well-known Speech Under Simulated and Actual Stress (SUSAS) [22] database, in which there are 11 pre-labelled emotional speech corpora. Details on the database are discussed in Section 3. Results show that the sinusoidal features of the eaQHM yield higher classification scores than those of the SM.

The rest of the paper is organized as follows. In Section 2 we will quickly review the analysis and synthesis steps of eaQHM. Section 3 presents the analysis parameters and the evaluation, both objective and subjective, of the eaQHM compared to SM. Section 4 discusses some preliminary classification issues, and finally, Section 5 discusses future perspectives and concludes the paper.

2. SHORT DESCRIPTION OF THE EAQHM-BASED ANALYSIS/SYNTHESIS SYSTEM

The speech signal is described as an AM-FM decomposition in the full-band (e.g. from 0 Hz up to the Nyquist frequency)

$$d(t) = \sum_{k=-K}^{K} A_k(t) e^{j\phi_k(t)} \tag{1}$$

where $A_k(t)$ is the instantaneous amplitude and $\phi_k(t)$ is the instantaneous phase of the k^{th} component, respectively. The instantaneous phase term is given by

$$\phi_k(t) = \phi_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} f_k(u) du \tag{2}$$

where $\phi_k(t_i)$ is the instantaneous phase value at the analysis time instant t_i, f_s is the sampling frequency, and $f_k(t)$ is the instantaneous frequency of the k^{th} component. The analysis part explains how to obtain the aforementioned parameters accurately. The analysis part is divided into two steps: an *initialization* step, where a first approximation of the speech signal is obtained under a harmonic assumption, and an *adaptation* step, where the parameters of the initialization step are iteratively refined.

2.1. Analysis - Initialization

A *continuous* f_0 estimation for all frames, noted by $\hat{f}_0$, is obtained at first, using the SWIPE pitch estimator [23] (although any pitch estimator can be used). The next step is to assume a full-band harmonicity to obtain a first estimate of the instantaneous amplitudes of all the harmonics. Using standard harmonic analysis [24], the parameters $|a_k(t_i)|, \phi_k(t_i)$ are obtained, where t_i is the i^{th} analysis time instant. Then, a first approximation of Eq. (1) can be obtained. Hence, $d(t)$ can be approximated by interpolating the $|a_k|$ and $\hat{f}_0$ values over successive analysis time instants t_i, resulting in

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{3}$$

where

$$\hat{A}_k(t) = |a_k(t)|, \quad \hat{\phi}_k(t_i) = \angle a_k(t_i) \tag{4}$$

and

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} (k\hat{f}_0(u) + \gamma(u)) du \tag{5}$$

where $\gamma(t)$ is a phase correction term to ensure phase coherence, as described in [18].

2.2. Analysis - Adaptation

In order to converge to quasi-harmonicity, the projection of the signal onto a set of amplitude and frequency varying basis functions is suggested in [19], using the parameters a_k and b_k of the Quasi-Harmonic Model (QHM) [25]. This yields the eaQHM model, which can be formulated in a single frame as:

$$d(t) = \left(\sum_{k=-L}^{L} \Big(a_k + tb_k\Big) \Big(\hat{A}_k(t) e^{j\hat{\phi}_k(t)}\Big) \right) w(t) \tag{6}$$

where $w(t)$ is the analysis window with support in $[-T, T]$, and $\hat{A}_k(t), \hat{\phi}_k(t)$ are defined as in Eqs. (4), (5). In this model, a_k, b_k are the complex amplitude and the complex slope of the k^{th} component, and $\hat{A}_k(t)$, $\hat{f}_k(t)$, $\hat{\phi}_k(t)$ are estimates of the instantaneous amplitude, frequency, and phase of the k^{th} component, respectively, from the initialization step. The a_k, b_k parameters are obtained via Least Squares [19]. The adaptation is completed by using the frequency correction mechanism first introduced in [25]. This mechanism provides a frequency correction $\hat{\eta}_k$, for each component. Hence, at the first adaptation, for the analysis time instant t_i, the instantaneous phases become

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} (\hat{f}_k(u) + \gamma(u)) du \tag{7}$$

where $\hat{f}_k(t) = k\hat{f}_0(t) + \hat{\eta}_k(t)$. Then, a Least Squares solution for the a_k, b_k using these refined frequencies (and phases) leads to a better estimation of the instantaneous amplitudes $\hat{A}_k(t) = |a_k(t)|$ and the $\hat{\eta}_k$ terms. By iteratively adding the $\hat{\eta}_k$ term of the current adaptation on the k^{th}-frequency track of the previous adaptation, the frequency tracks deviate from strict harmonicity and represent the underlying actual frequencies better. Finally, this adaptation scheme continues until a convergence criterion is met, which is related to the overall Signal-to-Reconstruction-Error Ratio (SRER) [26].

2.3. Synthesis

During synthesis, the k^{th} instantaneous amplitude track, $\hat{A}_k(t)$, is computed via linear interpolation of the successive estimates from the last adaptation step. The k^{th} instantaneous frequency track, $f_k(t)$, is computed via spline interpolation. As for the k^{th} instantaneous phase track, $\hat{\phi}_k(t)$, the non-parametric approach based on the integration of instantaneous frequency is followed, as is shown in the adaptation steps of the analysis. Finally, the speech signal can be approximated as:

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{8}$$

A block diagram of the algorithm is depicted in Figure 1.

3. ANALYSIS AND EVALUATION

In this section, the evaluation procedure is described, along with the dataset selection and the parameter estimation.

3.1. Objective Evaluation

At first, it is important to show that eaQHM can decompose high-quality running expressive speech signals into AM-FM components that represent the signal closer than SM. For this, a custom, small database of acted speech is used. This database consists of one male and one female subject, acting in four different speaking styles (*angry, sad, happy, neutral*), in a recording studio. A total number of

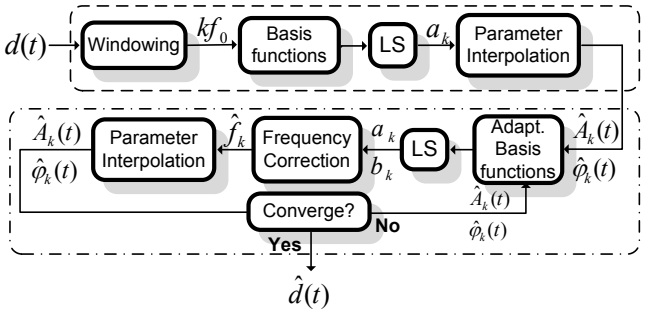

Fig. 1. *Block diagram of the eaQHM system. Dashed line includes the initialization (harmonic) part. Dot-dashed line includes the adaptation part.*

20 waveforms sampled at 16000 Hz are analyzed. All speech files in the database have been analyzed and resynthesized from their AM-FM components, and the corresponding SRER has been computed for each speech utterance. For this analysis, the window size was 30 ms for the SM and 3 local pitch periods for the eaQHM, both of Hamming type. A step size of 2.5 ms was selected for both models. The results are depicted in Table 1.

SRER Performance (Wideband Speech Database)				
Female Speaker				
	Speaking Styles			
Model	Angry	Happy	Neutral	Sad
SM	14.8 (1.36)	17.5 (3.0)	16.5 (1.36)	21.2 (1.64)
eaQHM	28.8 (1.24)	33.1 (1.81)	34.9 (2.23)	34.8 (3.60)
Male Speaker				
SM	17.0 (1.45)	14.3 (0.76)	16.0 (1.67)	16.5 (1.63)
eaQHM	35.7 (2.04)	31.6 (3.49)	33.3 (2.56)	33.1 (2.74)

Table 1. *Signal to Reconstruction Error Ratio values (dB) for both models on a small acted speech database. Mean and Standard Deviation are given.*

However, this database is not appropriate for classification purposes, since the containing data is too few. Another database will be used, named SUSAS (Speech Under Simulated and Actual Stress). The SUSAS database was developed in the 1990s and was the first emotional speech database ever created. It contains both actual and simulated stressed speech. In the simulated part, 9 U.S. English male speakers, of three main dialects (general USA, New England/Boston, and New York City accent), under different *simulated* stress conditions (*angry*, *clear*, *fast*, *lombard*, *loud*, *neutral*, *question*, *slow*, *soft*, and two conditions where the speaker was recorded during *medium and light activity*) have been recorded. Each speaking style corpus has 70 speech files per speaker, which consist of isolated, short communication words, such as *"hello", "break", "go"*, and *"destination"*. This amounts to about 1190 tokens per speaker, with a considerable subset of them being acoustically similar, such as (*six, fix*) and (*white, wide*). The simulated data in SUSAS database were sampled using a 16-bit A/D converter with sample rate of 8 kHz. Table 2 shows the mean and the standard deviation of SRER for all speakers, for most common speaking styles.

This clearly demonstrates the quality and the performance stability of the adaptive model compared to the SM on a large database of isolated words of different expressive speaking styles. It is interesting to note that both models appear to be very stable around a mean of about 16.6 and 32.5 dB, for the SM and the eaQHM respectively. Although the distribution of SRERs is wider in eaQHM-analysis, the mean is high enough to show that in almost all cases

SRER Performance (SUSAS)				
	Speaking Styles			
Model	Angry	Loud	Clear	Fast
SM	16.6 (3.06)	16.8 (3.01)	16.8 (3.06)	16.7 (3.03)
eaQHM	32.3 (5.61)	32.8 (5.59)	32.6 (5.62)	32.9 (5.58)
	Question	Soft	Neutral	Slow
SM	16.8 (3.00)	16.7 (3.05)	16.8 (3.01)	16.8 (3.05)
eaQHM	32.8 (5.57)	32.9 (5.61)	32.9 (5.58)	32.9 (5.60)

Table 2. *Signal to Reconstruction Error Ratio values (dB) for both models on the SUSAS database. Mean and Standard Deviation are given.*

the eaQHM manages to compactly capture most of the information present in the speech signal, for *all* speaking styles. Conclusively, it is evident that the adaptive model can handle word-isolated (i.e. SUSAS) and running expressive speech equally well.

3.2. Subjective Evaluation

For our subjective evaluation, a formal, on-line listening test was designed[1] using the small, high-quality database of emotional running speech. The listeners were asked to evaluate the overall quality of the resynthesized speech based on the two models. A total of 32 listeners participated in this test, and the results are depicted in Figure 2 along with the 95% confidence intervals. Please note that only 5 of them are familiar with signal processing. According to the preference test, almost all listeners noted eaQHM as being almost indistinguishable to the original one. It should be noted that the SUSAS

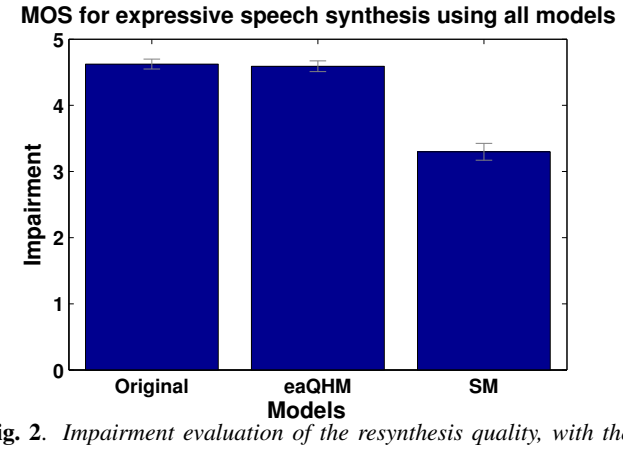

Fig. 2. *Impairment evaluation of the resynthesis quality, with the* 95% *confidence intervals.*

database was judged to perform poorly from a perceptual point of view due to the recording noise and the low sampling frequency. Informal listening tests showed that the eaQHM-based resynthesized speech samples were indistinguishable from the original ones, but this was the case for most samples obtained from the standard Sinusoidal Model as well. After careful listening, only a minority of waveforms demonstrated perceptual differences between the models but they were not enough in quantity to justify a listening test with this database. However, due to its pre-labelled data and its parallel corpora for each speaking style, this database was characterized as suitable for the classification task.

[1] http://www2.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.Exprtest

4. ON THE USE OF SINUSOIDAL PARAMETERS FOR EMOTION CLASSIFICATION

As already discussed, a discrimination between different emotional speaking styles is of great interest. Considering a sinusoidal analysis, it has been reported that amplitude and frequency values of the sinusoidal components can be used successfully to characterize the different expressive classes (emotions) in a speech signal [14]. Since the eaQHM can compute these parameters more accurately, it is not surprising that their discrimination properties among different speaking styles are similar or better than those reported in the literature for the standard SM. An example is presented in Figure 3, where the parameters of two speech samples (of the same word: *"No"*) from the SUSAS database pronounced with different emotional content (*angry, neutral*) are depicted.

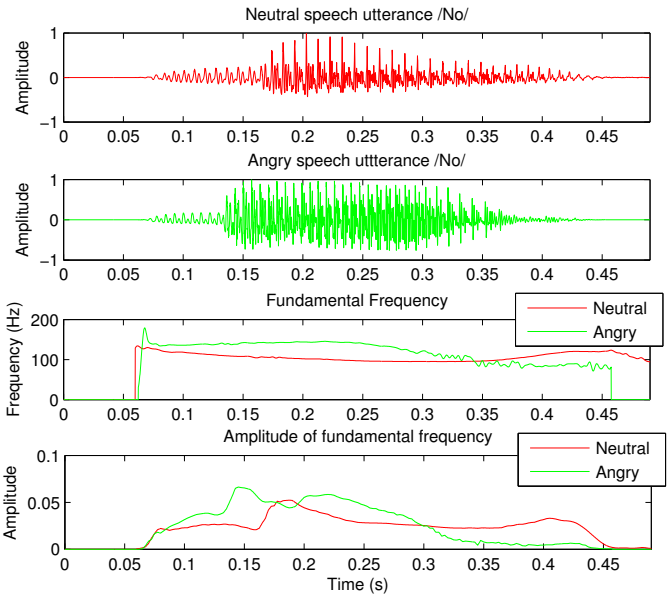

Fig. 3. *An example of analysis of emotional speech: First panel, neutral speech. Second panel, angry speech. Third panel, $f_0(t)$ tracks for each sample. Fourth panel, $A_0(t)$ tracks for each sample.*

Clearly, the amplitudes and frequencies of the fundamental are different in each case, and this is the case for other sinusoidal components as well. To verify this observation, a classification task based on a 128-bit Vector Quantizer (VQ) was designed as described in [14], using a subset corpus of the SUSAS, labelled as *Angry*, *Neutral*, *Soft*, and *Question*, containing a total number of 2520 waveforms (630 per emotion). A number of 756 waveforms were kept for testing (189 per emotion), while the rest were used for training. Both models were used for the analysis, at a frame rate of 2.5 ms, and the 10 highest amplitudes, along with their corresponding frequencies, were extracted from each analysis frame. The analysis window was set at 30 ms for the SM, and at 3 local pitch periods for the eaQHM. No distinction between voiced and unvoiced parts of speech was made. Two classification tasks were set, one using the amplitudes as features, and one using the frequencies. The Confusion Matrices for the amplitude-based classification are given in Tables 3 and 4, whereas for the frequency-based one are given in Tables 5 and 6.

From these results, the following observations can be made. In general, the parameters obtained from the eaQHM lead to better classification accuracy in all cases. Furthermore, the *angry* speaking style has the highest correct classification percentage for both models and both sets of features. The most difficult speaking style to classify correctly is the *question* one when the frequencies are used as features, and we can see that it is mostly confused with the *neutral* speaking style. On the other hand, the *question* speaking style has the lowest classification score when the amplitudes are used as features. A more robust classification system is expected to use a combination of parameters to make the classification decision. This is a task to be investigated in a future work.

		SM-based Classification - Amplitudes			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	72%	14%	3%	11%
	Neutral	4%	63%	18%	15%
	Soft	5%	30%	50%	15%
	Question	4%	22%	20%	55%

Table 3. *SM-based Confusion Table based on amplitudes for a* 128*-bit VQ classification between* 4 *emotions of the SUSAS database.*

		eaQHM-based Classification - Amplitudes			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	77%	14%	2%	7%
	Neutral	4%	64%	18%	14%
	Soft	3%	31%	56%	10%
	Question	6%	21%	13%	60%

Table 4. *eaQHM-based Confusion Table based on amplitudes for a* 128*-bit VQ classification between* 4 *emotions of the SUSAS database.*

		SM-based Classification - Frequencies			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	70%	7%	5%	18%
	Neutral	6%	38%	28%	27%
	Soft	3%	25%	59%	13%
	Question	18%	24%	25%	33%

Table 5. *SM-based Confusion Table based on frequencies for a* 128*-bit VQ classification between* 4 *emotions of the SUSAS database.*

		eaQHM-based Classification - Frequencies			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	71%	6%	7%	21%
	Neutral	6%	55%	24%	15%
	Soft	3%	13%	70%	14%
	Question	17%	18%	14%	50%

Table 6. *eaQHM-based Confusion Table based on frequencies for a* 128*-bit VQ classification between* 4 *emotions of the SUSAS database.*

5. CONCLUSIONS AND FUTURE WORK

In this work, we presented an application of an adaptive sinusoidal model, named eaQHM, on the problem of emotional speech analysis and classification. It was shown that different emotional speech styles can be effectively represented by the adaptivity mechanism of eaQHM, yielding very accurate AM-FM decomposition. This was demonstrated through resynthesis of the original speech signal from its AM-FM components and by evaluating the Signal-

to-Reconstruction Error (SRER). A formal listening test was designed to evaluate the perceptual quality of the resynthesized speech and showed that eaQHM-resynthesized emotional speech is indistinguishable from the original. Preliminary classification results showed that eaQHM-based classification achieves higher classification rates for a subset of the SUSAS database. Future work will focus on a more concrete classification scheme and an attempt to exploit phase-related features for classification purposes.

REFERENCES

[1] S. Bou-Ghazale and J. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 8, pp. 429 – 442, 2000.

[2] B. D.Womack and J. H. L. Hansen, "N-channel hidden markov models for combined stressed speech classification and recognition," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 7, pp. 668–676, 1999.

[3] J. H. L. Hansen and B. Womack, "Feature analysis and neural network based classification of speech under stress," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 4, pp. 307–313, 1996.

[4] J. H. L. Hansen, B. D. Womack, and L. M. Arsian, "A source generator based production model for environmental robustness in speech recognition," *In Proc. ICSLP*, pp. 1003 – 1006, 1994.

[5] B. D. Womack and J. H. L. Hansen, "Stress independent robust hmm speech recognition using neural network stress classification," *EUROSPEECH*, pp. 1999–2002, 1995.

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

[7] N. Amir and S. Ron, "Toward an automatic classification of emotions in speech," *Int. Conf. on Spoken Language Processing*, pp. 555–558, 1998.

[8] M. Bulut and S. Narayanan, "On the robustness of overall f0-only modifications to the perception of emotions in speech," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4547–4558, 2008.

[9] K. E. Cummings, M. A. Clements, and J. H. L. Hansen, "Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering," *In Proc. IEEE Southeastcon*, pp. 776–781, 1989.

[10] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, "Time and spectrum related variabilities in stressed speech under laboratory and real conditions," *Speech Communication*, vol. 20, pp. 111–130, 1996.

[11] A. Castellanos, J. M. Benedi, and F. Casacuberta, "An analysis of general acoustic - phonetic features for spanish speech produced with lombard effect," *Speech Communication*, vol. 20, pp. 23–36, 1996.

[12] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[13] L. T. Bosch, "Emotions, speech, and the ASR framework," *Speech Communication*, vol. 40, pp. 213–225, 2013.

[14] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 14, no. 3, pp. 737–746, 2006.

[15] M. W. Macon, D. D. J. Blumenthal, D. M. A. Clements, and D. R. M. Mersereau, "Applications of sinusoidal modeling to speech and audio signal processing," in *report in Georgia Institute of Technology*, 1993.

[16] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis based on a Sinusoidal Representation," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 34, pp. 744–754, 1986.

[17] C. Drioli, G. Tisato, P. Cosi, and F. Tesser, "Emotions and voice quality: Experiments with sinusoidal modeling," in *In Proceedings of VOQUAL03*, 2003, pp. 127–132.

[18] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, pp. 290–300, 2011.

[19] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An Extension of the Adaptive Quasi-Harmonic Model," in *Proc. IEEE ICASSP*, Kyoto, 2012.

[20] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085–2095, 2013.

[21] Y. Pantazis, "Adaptive AMFM signal decomposition with application to speech analysis," Ph.D. dissertation, Computer Science Department, University of Crete, 2010.

[22] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," *EUROSPEECH*, vol. 4, pp. 1743 – 1746, 1997.

[23] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, pp. 1628–1652, 2008.

[24] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, E.N.S.T - Paris, 1996.

[25] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the Properties of a Time-Varying Quasi-Harmonic Model of Speech," in *Interspeech*, Brisbane, 2008.

[26] Y. Pantazis, G. Tzedakis, O. Rosec, and Y. Stylianou, "Analysis/Synthesis of Speech based on an Adaptive Quasi-Harmonic plus Noise Model," in *Proc. IEEE ICASSP*, Dallas, Texas, USA, 2010.