Emotional Speech Classification using adaptive Sinusoidal Modelling

Theodora Yakoumaki[1,2], *George P. Kafentzis*[1], *Yannis Stylianou*[1]

[1]Department of Computer Science, University of Crete, Greece
[2]Institute of Computer Science, Foundation for Research and Technology Hellas, Greece
{yakumaki, kafentz, yannis}@csd.uoc.gr

Abstract

Automatic classification of emotional speech is a challenging task with applications in synthesis and recognition. In this paper, an adaptive sinusoidal model (aSM), called the *extended adaptive Quasi-Harmonic Model - eaQHM*, is applied on emotional speech analysis for classification purposes. The parameters of the model (amplitude and frequency) are used as features for the classification. Using a well known database of narrowband expressive speech (SUSAS), we develop two separate Vector Quantizers (VQ) for the classification, one for the amplitude and one for the frequency features. It is shown that the eaQHM can outperform the standard Sinusoidal Model in classification scores. However, single feature classification is inappropriate for higher-rate classification. Thus, we suggest a combined amplitude-frequency classification scheme, where the classification scores of each VQ are weighted and ranked, and the decision is made based on the minimum value of this ranking. Experiments show that the proposed scheme achieves higher performance when the features are obtained from eaQHM. Future work can be directed to different classifiers, such as HMMs or GMMs, and ultimately to emotional speech transformations and synthesis.

Index Terms: Adaptive quasi-harmonic model, Speech analysis, Emotional speech, Sinusoidal modelling, Emotion classification

1. Introduction

Speech produced from an emotionally charged speaker is defined as emotional (or stressed) speech. Speakers that feel *sad, angry, happy*, and *neutral* put a certain stress in their speech that is typically characterized as emotional. The emotional condition of the speaker may be revealed by the analysis of its speech, and such knowledge could be effective in emergency conditions, health care applications, and as a pre-processing step in recognition and classification systems. Moreover, applications in speech synthesis such as unit selection Text-to-Speech (TTS) synthesis, HMM-based synthesis, and speaker recognition and authentication applications could profit from such an analysis.

Acoustic analyses on speech produced under different emotional conditions reveals a great number of speech characteristics that vary according to the emotional state of the speaker. The variety of these features and their combination attributes to distinct emotional speech styles. The classification however, of emotional speech according to these features is a difficult task [1]. The use of Cepstral features and Linear Prediction coefficients (LP) in emotional speech analysis and classification was considered by Womack and Hansen [2, 3, 4, 5]. Improved results may be achieved with the Teager operator in contrast to the classification of emotional speech with the LP features [6]. In addition, it has been suggested that speaking styles can be identified by features that are connected with the pitch mean, variance and the intensity [7, 8]. Also, Cummings et al. [9] showed the variation of the glottal pulse shape among different emotional conditions. Frequency and time variabilities in emotional speech were presented by Ruiz et al. [10], whereas general acoustic-phonetic features were examined in Lombard speech by Castellanos et al. [11]. The intensity, spectral envelopes and duration in emotional speech were explored by Scherer [12] for speaker and speech recognition, whereas the usefulness of prosody for stressed speech recognition was discussed by Bosch [13]. The parameters of the Sinusoidal Model [14], namely amplitude, frequency, and phase were used as features to separate different speaking styles as suggested in [15]. For the recognition and/or classification of emotional speech, several classifiers have been suggested, such as Hidden Markov Models (HMM) [2, 15, 16, 17, 18, 19], Neural Networks (NN) [1, 3, 20, 21], Gaussian Mixture Models [22, 23], and Vector Quantization (VQ) [15, 24] using a variety of feature vectors.

Sinusoidal models (SMs) have not been involved in emotional speech analysis and/or classification until lately [15, 25]. Attempts using SMs are based on the features originated from the sinusoidal model parameters (amplitude, frequency, phase) over time. However, the estimation of these parameters is subjected to an important constraint; they are derived under the assumption of *local stationarity*, that is, the speech signal is assumed to be *stationary* inside the analysis window. Nonetheless, speech styles described as *fast* or *angry* may not hold this assumption. Recently, this problem has been handled by the adaptive Sinusoidal Model (aSMs) [26, 27], by projecting the signal onto a set of amplitude- and frequency-varying basis functions *inside* the analysis window. This way, sinusoidal parameters are more accurately estimated. Specifically, a phase adaptation of the basis function to the local characteristics of the signal is performed by the adaptive Quasi-Harmonic Model (aQHM) [26, 28], whereas both amplitude and phase adaptation is performed in the extended adaptive Quasi-Harmonic Model (eaQHM) [27]. Based on a full-band analysis-synthesis system described in [29], a first analysis on emotional speech has been conducted in [30], where eaQHM has demonstrated its ability to provide transparent resynthesized emotional speech and an average of 97% increase in Signal-to-Reconstruction-Error Ratio (SRER) compared to the standard Sinusoidal Model.

In this paper, the extended adaptive Quasi-Harmonic Model (eaQHM) is employed to classify emotional speech using its instantaneous parameters, namely the amplitude and frequency. The speech corpus for the analysis and classification is the well-known Speech Under Simulated and Actual Stress (SUSAS) [31] database, in which there are 11 pre-labelled emotional speech corpora. The classification is performed using a Vector Quantization scheme. Results show that the sinusoidal features of the eaQHM yield higher classification scores than those of the SM not only when the sinusoidal parameters are separately used but also when used in a combined scheme to achieve higher classification scores.

The rest of the paper is organized as follows. In Section 2 we will quickly review the analysis step of eaQHM. Section 3 presents the database of emotional speech and Section 4 presents the VQ-based emotion classification scheme. Finally, Section 5 suggests future perspectives and Section 6 concludes the paper.

2. Short description of the eaQHM-based analysis system

The eaQHM has been analytically presented in [27, 29]. Only the important points are considered here. The eaQHM is a powerful tool to accurately estimate the parameters of an AM-FM decomposed signal. Let us assume that 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. As depicted in Figure 1, 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 iterative *adaptation* step, where the model parameters are iteratively refined and diverge from strict harmonicity.

2.1. Analysis - Initialization

The recently proposed SWIPE pitch estimator [32] is applied every 1 ms on the speech signal to obtain a *continuous* f_0 estimation for all frames, noted by $\hat{f}_0$. Then, the full-band spectrum is sampled in integer multiples of the $\hat{f}_0$ in order to have a first estimate of the instantaneous amplitudes of all the harmonics. Standard frame-by-frame harmonic analysis [33] provides the parameters $|a_k(t_i)|, \phi_k(t_i)$, where t_i is the i^{th} analysis time instant. Then, $d(t)$ can be approximated by interpolating the $|a_k|$ and $\hat{f}_0$ values over successive analysis time instants t_i, resulting in an approximation of Eq. (1), where

$$\hat{A}_k(t) = |a_k(t)|, \quad \hat{\phi}_k(t_i) = \angle a_k(t_i) \tag{2}$$

and

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} (k\hat{f}_0(u) + \gamma(u)) du \tag{3}$$

are estimates of $A_k(t)$, $\phi_k(t)$, and $\gamma(t)$ is a phase correction term to ensure phase coherence, as described in [26].

2.2. Analysis - Adaptation

Since speech is not strictly harmonic, the representation of Eqs. (2), (3) is not entirely accurate. Thus, a refinement mechanism of the model parameters is applied; the projection of the signal onto a set of *amplitude and frequency varying* basis functions is suggested in [27], using the parameters a_k and b_k of the Quasi-Harmonic Model (QHM) [34]. This is the eaQHM model, which can be formulated in a single frame l as:

$$\hat{d}_l(t) = \left(\sum_{k=-L}^{L} \left(a_k^l + t b_k^l \right) \left(\hat{A}_k^l(t) e^{j\hat{\phi}_k^l(t)} \right) \right) w(t) \tag{4}$$

where $w(t)$ is the analysis window with support in $[-T, T]$, and $\hat{A}_k^l(t), \hat{\phi}_k^l(t)$ are estimates of the instantaneous amplitude and phase of the k^{th} component, respectively, extracted from the initialization step in Eqs. (2), (3) within the frame l. In this model, a_k, b_k are the complex amplitude and the complex slope of the k^{th} component.

The a_k, b_k parameters are obtained via Least Squares (LS) [27]. The reason to employ such a complex amplitude parameter scheme is that these parameters form a frequency correction mechanism, which was first introduced in [34]. This mechanism provides a frequency correction $\hat{\eta}_k$ for each sinusoidal component. So, adaptation can be defined as the successive frequency correction and instantaneous component (and thus, basis functions) re-estimation of the speech signal. Hence, at the first adaptation, at the analysis time instant t_i, the instantaneous phases become

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

where $\hat{f}_k(t) = k\hat{f}_0(t) + \hat{\eta}_k(t)$. Then, setting up exponential basis functions using Eq. (5) and solving for the a_k, b_k leads to a better estimation of the instantaneous amplitudes $\hat{A}_k(t) = |a_k(t)|$ and the $\hat{\eta}_k$ terms. The latter provides more accurate frequency components, and thus phase components. This iterative estimation of frequencies using the $\hat{\eta}_k$ term leads to a deviation from strict harmonicity and manages to represent the underlying signal more accurately. Finally, this adaptation scheme continues until a convergence criterion is met, which is related to the overall Signal-to-Reconstruction-Error Ratio (SRER) [26]. Then, the signal is reconstructed from its interpolated instantaneous parameters as in Eq. (1). A block diagram of the eaQHM is shown in Figure 1.

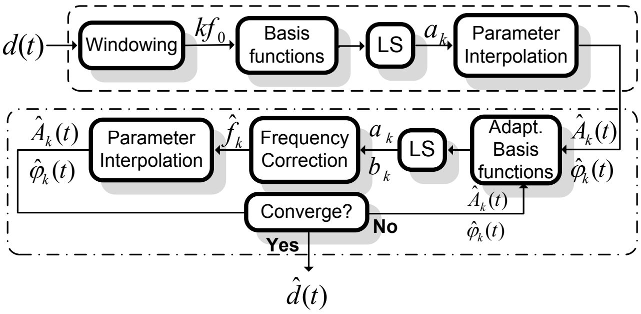

Figure 1: *Block diagram of the eaQHM system. Dashed line includes the initialization (harmonic) part. Dot-dashed line includes the adaptation part.*

3. Database Description

The Speech Under Simulated and Actual Stress (SUSAS) database was developed in the 1990s. 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 while performing medium and light physical activities) 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”, “point”, and “destination”. This sums to about 1190 tokens per speaker, with a considerable subset of them being acoustically similar, such as ("six", "fix") and ("white", "wide"). This fact, along with the small length of the utterances, makes the database difficult enough for several applications, such as speech recognition and emotion classification. The simulated data in SUSAS database were sampled using a 16-bit A/D converter with sample rate of 8 kHz.

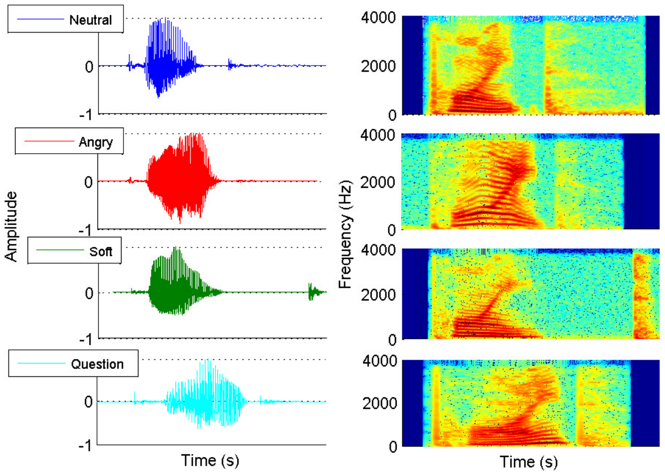

Figure 2: *An example of emotional speaking styles, in time and frequency: First panel, neutral. Second panel, angry. Third panel, soft. Fourth panel, question. The word "Point" is depicted in this example.*

4. VQ-based 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 different expressive classes (emotions) in a speech signal [15]. 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 of a single word ("point") in four different emotions is depicted in Fig. 2, along with the corresponding spectrograms that partly reveal their differences. The signals are aligned according to the stop consonant /p/. It can be seen that these differences appear in amplitude strength, frequency variations, energy distributions, formant positioning, timings, duration of vowels and consonants, etc. Sinusoidal modeling can capture some of these differences in the form of AM-FM components [15]. Due to its adaptive processing, we propose that eaQHM can provide parameters that are highly accurate, which makes them more suitable for an emotion classification task than the same parameters obtained from a standard SM.

4.1. Feature Extraction

To evaluate our suggestion, a classification task based on a 128-bit Vector Quantizer (VQ) was designed using a subset corpus of the SUSAS, labelled as *Angry*, *Neutral*, *Soft*, and *Question*. A total number of 2520 waveforms (630 per emotion) were used. A number of 756 waveforms were kept for testing (189 per emotion), while the rest were used for training. All discrete-time waveforms were normalized to unit energy, as in

$$x[n] = \frac{x[n]}{\sqrt{\sum_{n=0}^{L-1} x^2[n]}} \tag{6}$$

where L is the signal length in samples. Both models used an analysis frame rate of 2.5 ms. The 10 strongest components of the magnitude spectrum of the FFT and the 10 highest sinusoidal amplitudes provided by the LS, 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 in this work.

4.2. Classification - Single Feature

At first, two classification tasks were set, each one using different features (amplitudes and frequencies). Having M spectral vectors $\mathbf{x_i}$ containing the selected features (amplitudes or frequencies), the data matrix $\mathbf{X}$ is created as $\mathbf{X} = [\mathbf{x_1} \quad \mathbf{x_2} \quad \cdots \quad \mathbf{x_M}]$. The codebooks are then designed based on the minimization of the Average Distortion (AD) between the training vectors and the codebook vectors in matrix $\mathbf{Y}$, where $\mathbf{Y} = [\mathbf{y_1} \quad \mathbf{y_2} \quad \cdots \quad \mathbf{y_C}]$, and C is the codebook size. The AD is defined as

$$AD = \frac{1}{C} \sum_{k=1}^{C} \min_{\mathbf{y_i} \in \mathbf{Y}} d^2(\mathbf{x_k}, \mathbf{y_i}) \tag{7}$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean Distance (ED) between vectors $\mathbf{x}$ and $\mathbf{y}$. For each of the four emotions mentioned earlier, a codebook was designed using the LBG algorithm [35]. The emotion is recognized by the minimum average distortion. The Confusion Matrix for the amplitude-based classification is given in Table 1, whereas for the corresponding frequency-based one is given in Table 2. It can be seen that in both cases the *angry* speaking style stands out of the rest of speaking styles. This is expected since this speaking style is very different than the others in terms of amplitude and frequency distributions [15].

		VQ Classification in % - Amplitudes			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	77(72)	14(14)	2(3)	7(11)
	Neutral	4(4)	64(63)	18(18)	14(15)
	Soft	3(5)	31(30)	56(50)	10(15)
	Question	6(4)	21(22)	13(20)	60(55)

Table 1: *Classification score (%) for four emotions of the SUSAS database, using amplitude features extracted from eaAQHM and SM (in parenthesis).*

		VQ Classification in %- Frequencies			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	71(70)	6(6)	7(5)	21(18)
	Neutral	6(6)	55(38)	24(28)	15(27)
	Soft	3(3)	13(25)	65(59)	14(13)
	Question	17(18)	18(24)	14(25)	50(33)

Table 2: *Classification score (%) for four emotions of the SUSAS database, using frequency features extracted from eaAQHM and SM (in parenthesis).*

In general, the parameters obtained from the eaQHM lead to better classification scores in all cases. Furthermore, the *angry* speaking style has the highest correct classification percentage for both models and both sets of features. The *question* speaking style is the most difficult one to correctly classify 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 *soft* speaking style has the lowest classification score when the amplitudes are used as features.

4.3. Classification - Combined Features

Since single-feature based classification leads to low classification scores, a combined classification scheme is suggested. The ADs obtained from amplitude and frequency based VQs are normalized by the highest corresponding AD. Then, the ADs of the corresponding emotions are added. Finally, the emotion with the minimum sum of ADs is selected as the recognized emotion. This way, when the VQs have decided differently, the VQ which is more "confident" in its decision (the minimum AD is far less than other ADs) can influence the final outcome. Figure 3 illustrates the proposed scheme.

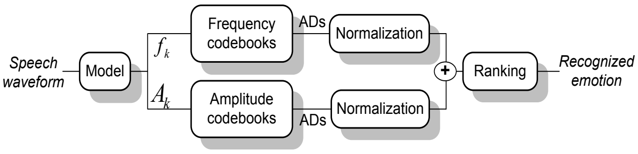

Figure 3: *The proposed classification scheme based on the combination of features.* A_k *and* f_k *denote the instantaneous amplitude and frequency components, and* ADs *denote the average distortion measures.*

Table 3 presents the corresponding classification scores for eaQHM and SM using the proposed scheme. Using this scheme, on average, the eaQHM correctly classifies 65% of the utterances in the database, whereas the SM reaches 54%. Apparently, not all speaking styles were favoured by this combined scheme. Mostly the *angry* and the *question* speaking style achieved significant increase of their classification rates in both models. While the *angry* speaking style already had a relatively high percentage, the *question* speaking style has interestingly increased its correct classification score. However, the *soft* and *neutral* speaking style did not significantly change their percentages. This suggests that a weighted sum of the ADs before ranking may be more appropriate.

		VQ combined classification in %			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	83(77)	5(5)	1(5)	11(13)
	Neutral	15(4)	58(48)	12(24)	15(24)
	Soft	10(2)	18(29)	56(54)	16(15)
	Question	20(17)	6(24)	11(21)	63(38)

Table 3: *eaQHM and SM based Confusion Table in % based on amplitudes and frequencies for a* 128*-bit VQ classification between* 4 *emotions of the normalized SUSAS database.*

5. Discussion and Perspectives

In this work, we attempted to perform emotion classification from speech signals using instantaneous parameters of sinusoidal models. Although the database in hand contains short, isolated words with similar perceptual content, and this makes recognition and classification results rather difficult, results are encouraging. However there is room for improvement.

First of all, the use of phase information could be exploited in combination with amplitudes and frequencies. In [15], the number of *phase reversals* is suggested as a feature. However, a more intuitive measure could be suggested. In [36], the notion of relative phase shift (RPS) is revisited and a phase structure is shown to be revealed through RPS. It would be interesting to examine if there are different patterns in RPS structures that can help discriminate emotional content in speech, combined with the standard amplitude and frequency features.

Secondly, sinusoidal amplitudes provide an implicit information about the spectral envelope, and they have been shown to be important in emotion recognition [2, 3, 4]. Nevertheless, when considering only a part of the full-band, such as the 10 highest spectral peaks, a significant part of the spectrum is not taken into account. The inclusion of that part may contribute to better recognition percentages. Moreover, higher frequency components were suggested to be disregarded in sinusoidal model-based emotion classification as inappropriate for the task [15]. However, the aSMs are able to follow the dynamics of speech in the upper bands, and thus to reveal the spectral details that are blurred due to the time-frequency trade-off of the FFT-based estimation.

Furthermore, vowels have received increasing attention when it comes to emotion recognition, however consonants are shown to be important as well (see for example [37]). Since our model is full-band and models both voiced and unvoiced parts of speech using AM-FM components, it would be interesting to examine whether there is any useful information embedded in the sinusoidal representation of consonants that is able to distinguish emotions. Finally, different classifiers can be used, such as HMMs, SVMs, or GMMs, for a more efficient classification.

6. Conclusions and Future Work

In this paper, we presented an application of an adaptive sinusoidal model, named eaQHM, on the problem of emotional speech classification and compared it to the standard Sinusoidal Model. The instantaneous amplitude and frequency were used as features for the classification. Results showed that a Vector Quantization classification based on eaQHM achieves higher classification scores for a subset of the SUSAS database, both on single-feature classification based on the sinusoidal parameters and on their combination. Future work will focus on different classifiers, phase parameter exploitation, and transforming neutral speech into emotional.

7. References

[1] S. Bou-Ghazale and J. Hansen, "A comparative study of traditional and newly proposed features for recognitionof speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 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 environmen-

tal 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] 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.

[15] 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.

[16] D. A. Cairns and J. Hansen, "Nonlinear analysis and classification of speech under stressed condition," *J. Acoust. Soc. Am.*, pp. 3392–3400, 1994.

[17] N. Nogueiras, A. Moreno, A. Bonafonte, and J. Marino, "Speech emotion recognition using Hidden Markov Models," in *EUROSPEECH*, 2001, pp. 2679–2682.

[18] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," in *EUROSPEECH*, 2003, pp. 125–128.

[19] T. Nwe, S. Foo, and L. D. Silva, "Speech emotion recognition using Hidden Markov Models," *Speech Communication*, vol. 41, pp. 603–623, 2003.

[20] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing & Applications*, vol. 9, pp. 290–296, 2000.

[21] M. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," in *Proc. of the International Symposium on Circuits and Systems*, vol. 2, 2004, pp. II–181–4 Vol.2.

[22] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Interspeech*, 2005, pp. 493–496.

[23] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian Mixture Vector autoregressive models," in *Proc. IEEE ICASSP*, 2007, pp. 957–960.

[24] P. Khanna and M. S. Kumar, "Application of vector quantization in emotion recognition from human speech," in *Information Intelligence, Systems, Technology and Management*. Springer Berlin Heidelberg, 2011, vol. 141, pp. 118–125.

[25] 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.

[26] 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.

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

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

[29] G. P. Kafentzis, O. Rosec, and Y. Stylianou, "Robust full-band adaptive sinusoidal analysis and synthesis of speech," in *Proc. IEEE ICASSP*, 2014, in Press. [Online]. Available: https://www.researchgate.net/publication/260034866

[30] G. P. Kafentzis, T. Yakoumaki, A. Mouchtaris, and Y. Stylianou, "Analysis of emotional speech using an adaptive sinusoidal model," in *EUSIPCO*, 2014, submitted for publication.

[31] 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.

[32] 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.

[33] 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.

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

[35] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[36] I. Saratxaga, I. Hernáez, D. Erro, E. Navas, and J. Sánchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 7, 2009.

[37] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613–625, 2010.