HIGH-RESOLUTION SINUSOIDAL MODELING OF UNVOICED SPEECH

George P. Kafentzis and Yannis Stylianou

Multimedia Informatics Lab
Department of Computer Science
University of Crete, Greece

ABSTRACT

In this paper, a recently proposed high-resolution Sinusoidal Model, dubbed the extended adaptive Quasi-Harmonic Model (eaQHM), is applied on modeling unvoiced speech sounds. Unvoiced speech sounds are parts of speech that are highly non-stationary in the time-frequency plane. Standard sinusoidal models fail to model them accurately and efficiently, thus introducing artefacts, while the reconstructed signals do not attain the quality and naturalness of the originals. Motivated by recently proposed non-stationary transforms, such as the Fan-Chirp Transform (FChT), eaQHM is tested to confront these effects and it is shown that highly accurate, artefact-free representations of unvoiced sounds are possible using the non-stationary properties of the model. Experiments on databases of unvoiced sounds show that, on average, eaQHM improves the Signal to Reconstruction Error Ratio (SRER) obtained by the standard Sinusoidal Model (SM) by 93%. Moreover, modeling superiority is also supported via informal listening tests with two other models, namely the SM and the well-known STRAIGHT method.

***Index Terms*—** Sinusoidal Model, extended Adaptive Quasi-Harmonic Model, Speech Analysis, Unvoiced Speech

1. INTRODUCTION

Representing speech in an intuitive and compact way is a challenging problem that has gained significant attention since the start of the digital computer era. Many state-of-the-art systems include the so-called Sinusoidal Model (SM) [1] for modeling the speech spectral content, exploiting its inherent ability in accurately capturing the quasi-periodic phenomena that typically occur in speech signals. The SM treats unvoiced parts of speech the same way as voiced ones, based on the principle that the periodogram peaks are close enough to satisfy the requirements imposed by the Karhunen-Loeve expansion [2]. Furthermore, more sophisticated models decompose speech into deterministic and stochastic components and can provide high-quality representations of a given speech signal, well-suitable for applications such as transformations [3, 4, 5], conversion [4, 6], and speech synthesis [7, 8]. The success of sinusoidal models led to a number of refinements, as (for example) in spectral estimation [9, 10, 11] and unvoiced speech modeling [4, 12, 13].

When discussing about unvoiced speech, one can understand that it consists of signals whose nature is either noise-like (called *fricatives*), silence-like followed by a sharp attack (called *stops*), or a combination (called *affricates*). A *stop* sound is produced with complete closure of the articulators involved, so that the stream of air cannot escape through the mouth. *Voiced stops* are produced with vibrating vocal folds whereas in *voiceless stops* vocal folds are apart. A *fricative* is produced with close approximation of the two articulators, so that the stream of air is partially obstructed and turbulent airflow is produced. Finally, an *affricate* is a stop, followed by a fricative sound.

Although unvoiced speech used to be less popular in applications than voiced speech, there are numerous recent works that utilize a representation of unvoiced speech. In [14], emotion detection and classification of speech is presented, using a standard sinusoidal representation of voiced and unvoiced speech, utilizing the sinusoidal parameters as features for the classifiers. Moreover in [15], a very similar approach is followed for speech emotion recognition, taking sinusoidal parameters and their first- and second-order differences into account. Unvoiced speech is also included in this work, as well as elsewhere [16, 17]. Finally, applications such as time- and pitch-scaling can benefit from a sinusoidal representation of unvoiced speech [18, 19].

From a technical point of view, a sinusoidal representation of unvoiced speech is appealing for two main reasons: (1) locating the voicing boundaries when separating voiced from unvoiced speech is not an easy task, and (2) separate manipulation of deterministic and stochastic components increases the risk that listeners perceive them as separately processed. However, it is questionable how and why sinusoids are appropriate when representing these consonants. When dealing with unvoiced speech, approaches that assume stationarity inside the analysis window suffer from artefacts, such as the so-called pre-echo effect [20, 21], that is inherent in the Fourier Transform mostly used in these methods, and from reduced intelligibility due to the misrepresentation of the stochastic content by stationary sinusoids. The main reason behind these problems is that unvoiced speech is represented by stationary sinusoids *inside* an analysis window. Thus, in the literature, many alternatives include the use of short analysis windows combined with multi-resolution techniques when unvoiced sounds are detected as in [21], but this does not alleviate neither the pre-echo effect (in stop sounds), nor the reconstruction quality (in other unvoiced sounds). Ultimately, copy strategies [22], transform coding [21] [23], or modulated noise [4, 12, 22] are used instead.

A first step towards modeling unvoiced speech has been presented in [20], where voiceless (and their corresponding voiced) stop sounds were very efficiently modeled using an adaptive Sinusoidal Model, dubbed extended adaptive Quasi-Harmonic Model (eaQHM) [24], as high-resolution, non-stationary, time-varying sinusoids. It has been shown that these models can adapt to the analyzed signal better than typical sinusoidal representations, therefore achieving high reconstruction quality, as measured by the Signal-to-Reconstruction-Error Ratio (SRER) [24, 25]. Experiments showed that eaQHM provides a nearly pre-echo-free representation of stop

sounds, without the necessity of using very short analysis window lengths for these sounds, neither the use of a transient detector as in [21].

To the direction of fricatives and affricates, let us examine a sample more closely using the Fast Fourier Transform (FFT) and the recently proposed Fan-Chirp Transform (FChT) [26, 27]. In Figure 1, a fricative /s/ is depicted, along with the corresponding spectrograms based on the FFT and the FChT. Although in the FChT there are not any prominent time-frequency tracks that can justify a sinusoidal model framework, intuitively, an adaptive decomposition of unvoiced speech should attempt to locate "optimal" frequency tracks that collectively minimize the mean-square error inside the frame. These "optimal" frequency tracks become more discernible in the FChT-based spectrogram, whereas in the DFT-based spectrogram severe blurring still exists.

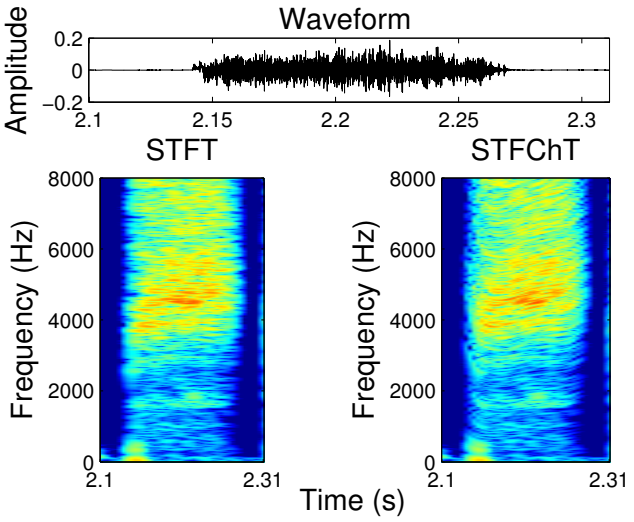

Fig. 1. *Spectral analysis of unvoiced speech. Top: Unvoiced speech waveform, Bottom left: FFT-based spectrogram slice of the corresponding waveform. Bottom right: FChT-based spectrogram slice of the corresponding waveform. Horizontal axis is time in seconds in all figures.*

In this paper, eaQHM is applied on the problem of modeling unvoiced speech, and more specifically, fricative and affricate sounds. We will show how adaptivity (1) can compensate the analysis problems of such sounds and (2) is capable of accurately representing them as AM-FM components. Experiments are conducted on a large database of more than 400 isolated sounds, and SRER measures are presented and discussed. Finally, subjective listening tests reveal that adaptive sinusoids perceptually outperform the baseline model (SM) and a state-of-the-art representation (STRAIGHT) [28].

The rest of the paper is organized as follows. In Section 2, we quickly review adaptive Sinusoidal Modeling, and especially the eaQHM. Section 3 presents a fricative as a case study, and the limitations of classic sinusoidal modeling versus adaptive modeling are revealed. Section 4 compares two well-know sinusoidal-based speech representations (standard Sinusoidal Model and STRAIGHT) with the eaQHM in modeling a large speech database of unvoiced sounds. SRER measures are provided and the relative performance is discussed. Section 5 presents the results of a formal listening test based on sinusoidal resynthesis of unvoiced speech. Finally, Section 6 concludes the paper.

2. ADAPTIVE SINUSOIDAL MODELING

The aSMs utilize the Least-Squares minimization criterion to estimate the parameters. The *adaptive* term is justified by successive refinements of the model basis functions based on instantaneous parameter re-estimation.

In general, an aSM can be described as

$$x(t) = \left(\sum_{k=-K}^{K} C_k(t)\psi_k(t) \right) w(t) \tag{1}$$

where $\psi_k(t)$ denotes the set of basis functions, $C_k(t)$ denotes the (complex) amplitude term of the model, $2K+1$ is the number of exponentials (hence, $K+1$ sinusoids), and finally $w(t)$ is the analysis window with support in $[-T, T]$.

Using this notation, in conventional sinusoidal models (including the SM, the Harmonic Model (HM) [4], the Quasi-Harmonic Model (QHM) [29], and others), the set of basis functions $\psi_k(t)$ in the analysis part is stationary in frequency and in amplitude. For example, the basis functions in the SM are in the form of

$$\psi_k^{SM}(t) = 1 \cdot e^{j2\pi f_k t}, \qquad C_k^{SM}(t) = a_k \tag{2}$$

where the amplitudes and frequencies of the basis functions are constant (in other words, stationary) inside the analysis window (1 and f_k, respectively). On the contrary, eaQHM does not share this assumption.

Specifically, eaQHM projects a signal segment onto a set of non-parametric, time-varying basis functions with instantaneous amplitudes and phases that are adapted to the local characteristics of the underlying signal [24]:

$$\psi_k^{eaQHM}(t) = \hat{A}_k(t)e^{j\hat{\Phi}_k(t)}, \quad C_k^{eaQHM}(t) = (a_k + tb_k) \tag{3}$$

where a_k and b_k are the complex amplitude and the complex slope of the model respectively, and $\hat{A}_k(t)$, $\hat{\Phi}_k(t)$ are functions of the instantaneous amplitude and phase of the signal, given by

$$\hat{A}_k(t) = \frac{|a_k(t)|}{|a_k(0)|} \quad , \quad \hat{\Phi}_k(t) = \hat{\phi}_k(t) - \hat{\phi}_k(0) \tag{4}$$

Both instantaneous parameters are obtained from an initialization step (a preliminary estimation and interpolation of the instantaneous parameters). Clearly, $\psi_k^{eaQHM}(t)$ define basis functions that vary inside the analysis window.

The instantaneous phase $\hat{\phi}_k(t)$ is computed using a frequency integration scheme [25], although cubic phase interpolation could be used as well [1]. The instantaneous amplitude $|a_k(t)|$ is estimated via linear interpolation, while $f_k(t)$ is estimated via spline interpolation. The eaQHM is actually a parameter-refinement mechanism, thus it requires an initialization, as already mentioned. For this purpose, any AM-FM decomposition algorithm can be used, but in most of the previous works concerning the eaQHM [24, 30], the Harmonic Model (HM) [4] or the Quasi-Harmonic Model [29] is used.

Considering that a preliminary estimation of the instantaneous components $|\hat{a}_k(t)|$ and $\hat{\phi}_k(t)$ of the signal is available, the estimation of the unknown parameters of eaQHM is similar to that of the

Harmonic Model or the Quasi-Harmonic Model, using the Least-Squares minimization method. However, the basis functions are both non-parametric and non-stationary. Parameters $\hat{A}_k(t)$ and $\hat{\Phi}_k(t)$ are iteratively refined using a_k and b_k, forming a frequency correction term $\hat{\eta}_k$ for each sinusoid, first introduced in [29], Applying the $\hat{\eta}_k$ on each frequency track, interpolating the instantaneous parameters over successive frames and restructuring the basis functions leads to more accurate model parameter estimation. These form a new frequency mismatch correction, $\tilde{\eta}_k$. This way, the loop goes on until the instantaneous parameters yield a close representation of the underlying signal, according to a Signal-to-Reconstruction-Error Ratio (SRER) based criterion [24, 20]. Finally, the signal is reconstructed from its AM-FM components as

$$s(t) = \sum_{k=-K}^{K} |\hat{a}_k(t)| e^{j\hat{\phi}_k(t)} \tag{5}$$

where $\hat{\phi}_k(t)$ is formed by a frequency integration scheme [25].

After applying the eaQHM for a number of adaptations, the instantaneous parameters are interpolated over successive frames and the overall signal is synthesized as in Eq. (5). It should be emphasized that the standard SM and eaQHM end up in the *same* number of parameters per time instant t_i for resynthesis (three parameters per frame, namely the amplitude $|a_k(t_i)|$, the frequency $f_k(t_i)$, and the phase $\phi_k(t_i)$). For more insight on eaQHM and the adaptation algorithm, please refer to [24].

3. ADAPTIVE SINUSOIDAL MODELLING OF UNVOICED SPEECH

As a reminder, fricatives are consonants produced by forcing air through a narrow passage made by placing two articulators close together, while affricates consist of a stop sound, followed by a fricative. For modelling such sounds, a similar strategy as for stop sounds [20] is followed for their analysis. A test case of a fricative /s/ is depicted in Figure 3, where the reconstructed signals from eaQHM (Fig. 3, right) and SM (Fig. 3, left) are presented, along with their corresponding residuals. As expected, that the adaptive model will finetune its local parameters to the local energy maxima of the spectrum, through its inherent frequency correction mechanism. The basis functions of the successive adaptation steps will be formed by the corrected parameters, thus giving AM-FM components that come more and more closer to the spectral characteristics of the waveform.

In technical details, the signal is sampled at $F_s = 16$ kHz, and a low initial frequency value such as 80 Hz, which results in frequency values of $80k$ Hz, $k = -100, \cdots, 100$, is chosen for both models. Hence, the frequencies cover the full-band of the spectrum. The frame rate is set to 1 sample and the analysis window is three times the local pitch period, that is $3/80$ seconds, and is of Hamming type. Same settings are applied to the Sinusoidal Model. The SRER performance of eaQHM was found to be 33.03 dB, over the 8.86 dB of the standard SM. Clearly, eaQHM outperforms SM by more than 400% in this test case. Thus, eaQHM seems to be promising for modeling unvoiced sounds. Figure 2 shows how the SRER evolves over the adaptation number, starting from 14.2 dB without any adaptation - performing simple Least Squares minimization on purely harmonic basis functions - and reaching up to 33.03 dB on the 5^{th} adaptation. The initial harmonic grid does not fully capture the present spectral energy, but successive adaptations locally finetune the frequencies, resulting in a remarkably better spectral representation of the sound.

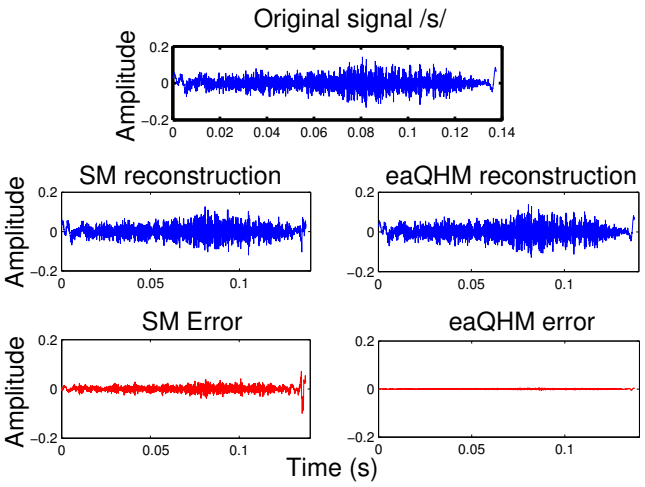

Fig. 2. *Estimated waveforms for a fricative sound /s/. Upper panel: Original signal. Middle panel: SM (left) reconstruction and eaQHM (right) reconstruction. Lower panel: SM (left) and eaQHM (right) reconstruction error.*

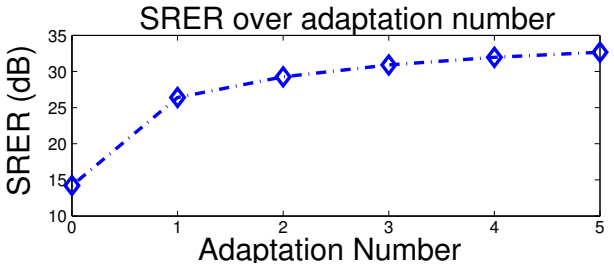

Fig. 3. *SRER evolution over adaptation number for eaQHM for the test case signal /s/. Adaptation number* 0 *stands for no adaptation (stationary basis functions).*

4. OBJECTIVE EVALUATION

To validate and extend our assumption, 485 voiceless fricatives and affricates (and their corresponding voiced ones, for comparison purposes) have been automatically extracted from speech in English uttered by a male and a female subject and analyzed using both the SM and eaQHM. Voiced fricatives include /v/, /ð/, /s/, and /ʃ/, while unvoiced ones are /f/, /θ/, /z/, and /ʒ/. Affricates include /tʃ/ and /dʒ/. The number of samples extracted from the male speaker was almost the same as those from the female speaker. The frame rate of 1 sample used in the previous section is not realistic for applications. Thus, the frame rates selected are 1 ms, 2 ms, and 4 ms. Parameters other than the frame rate remain the same as in the previous section. Table 1 presents the results per speech sound, in terms of mean value of SRER. It is

As it can be observed from Table 1, the performance of the adaptive model sustains in high reconstruction levels, even with a frame

Validation for Unvoiced Speech											
Signal to Reconstruction Error Ratio (dB)											
		Fricatives								Affricates	
Step	Model	/v/	/ð/	/s/	/ʃ/	/f/	/θ/	/z/	/ʒ/	/tʃ/	/dʒ/
1 ms	SM	14.7	13.2	13.9	11.3	12.7	15.1	17.5	17.3	11.3	11.5
	eaQHM	26.4	25.6	24.1	26.4	25.8	24.3	29.5	28.9	25.1	24.8
2 ms	SM	13.1	11.3	12.1	10.4	10.2	14.7	15.9	15.2	11.0	10.9
	eaQHM	23.5	23.1	22.6	24.7	23.5	22.6	28.6	27.8	23.1	23.8
4 ms	SM	12.2	10.6	11.2	9.6	9.7	8.9	13.3	13.7	9.3	10.2
	eaQHM	22.4	22.2	21.9	23.1	22.6	21.7	27.5	27.1	21.5	22.3

Table 1. *Signal to Reconstruction Error Ratio values (dB) for all models on a large database of fricatives and affricates. Step denotes the analysis frame rate.*

rate up to 4 ms. The mean standard deviation per model is: 3.4 dB (SM) and 4.1 dB (eaQHM). No significant variations in standard deviation were observed across different sounds. Experiments with higher frame rates were performed as well, such as 5 and 10 ms, that showed an average decrease of 3.9 and 6.5 dB respectively, compared to the 4 ms case, for all sounds for eaQHM. The SM showed an average decrease of 4.1 and 7.8 dB compared to the 4 ms case. Therefore it is suggested, as a rule of a thumb, the use of as low frame rate as possible to attain a high enough perceptual and reconstruction quality. The average number of adaptations required for the convergence is found to be 3.8, 4.1, and 4.7 for eaQHM, for step sizes of 1, 2, and 4 ms, for all sounds.

5. SUBJECTIVE EVALUATION

Since isolated unvoiced sounds are hard to be subjectively evaluated mainly due to their short duration, the performance of the algorithms are tested on the basis of full speech waveform reconstruction using eaQHM as a full signal model, as described in [30]. The goal of the listening test was not only to evaluate the perceived quality of the resynthesized unvoiced speech, but to reveal the advantages of having a single deterministic model for *all* parts of speech. Listeners were asked to evaluate the similarity between each one of 28 recordings of short words and their corresponding reconstruction using SM, STRAIGHT, and eaQHM. Also, the listeners were requested to absolutely focus on the quality of unvoiced speech, compared to the original. The waveforms were sampled at $F_s = 16$ kHz. For the analysis of sinusoidal models, the window length is 3 times the local pitch period, obtained from the well-known SWIPE pitch estimator [31]. The window type is Hamming for both models, and the frame rate is 1 ms (best performance according to Table 1) for all three models. For synthesis, parameter interpolation is selected for both sinusoidal models. For STRAIGHT, the default parameters are used. In total, 300 and 2051 parameters per frame are required for resynthesis using both sinusoidal models and STRAIGHT, respectively. 12 listeners participated in the test using only high-quality headphones in a quiet laboratory environment, and the Mean Opinion Scores (MOS) are presented in Figure 4. Apparently, eaQHM provides transparent perceived quality of unvoiced speech, compared to the stationary sinusoidal approach of the SM and the aperiodicity component which models non-deterministic parts of speech of the STRAIGHT method.

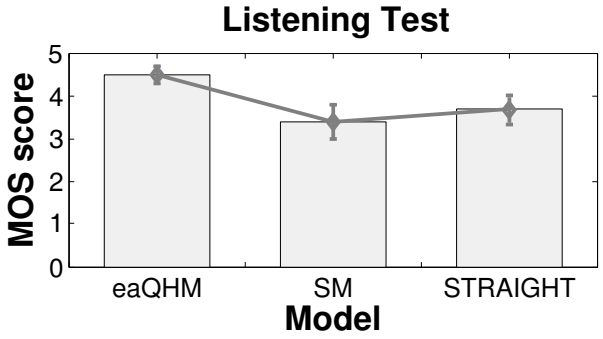

Fig. 4. *Listening Test based on Mean Opinion Score (MOS), along with the* 95% *confidence intervals.*

6. CONCLUSIONS

In this paper, high-resolution modeling of unvoiced speech sounds is presented and addressed via the extended adaptive Quasi-Harmonic Model. It is shown that local adaptation of the analysis parameters results in AM-FM components that are able to decompose and reconstruct unvoiced sounds effectively. SRER measures validate the latter for different unvoiced speech categories and different frame rates. It is found that eaQHM gives an average of 93% higher SRER values compared to the standard Sinusoidal Model. Listening tests also verified the transparency of the reconstruction quality. The latter is important to support the transition from hybrid speech models to full-band ones that operate on the full length of the speech signal, without any quality degradation, and thus providing a uniform and highly accurate representation of speech as high resolution AM-FM components. Future work will focus mostly on speech transformations, since the preservation of the modeled unvoiced parts under modification (pitch and time scale) is promising.

7. REFERENCES

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

[2] H. Van Trees, *Detection, Estimation, and Modulation Theory: Part I*, Wiley, New York, 1968.

[3] J. Laroche Y. Stylianou and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proceedings of EUROSPEECH*, 1995.

[4] Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, E.N.S.T - Paris, 1996.

[5] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 389–406, 1997.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[7] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 21–29, 2001.

[8] M. Macon, *Speech Synthesis Based on Sinusoidal Modeling*, Ph.D. thesis, Georgia Institute of Technology, 1996.

[9] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT–a subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1340–1342, 1986.

[10] S. Van Huffel, H. Park, and J.B. Rosen, "Formulation and solution of structured total least norm problems for parameter estimation," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2464–2474, 1996.

[11] R. B. Dunn and T. F. Quatieri, "Sinewave Analysis/Synthesis Based on the Fan-Chirp Transform," *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October, 2007.

[12] X. Serra, *A System for Sound Analysis, Transformation, Synthsis based on a Determistic plus Stochastic Decomposition*, Ph.D. thesis, Stanford University, 1989.

[13] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," in *IEEE Transactions on Speech and Audio Processing*, 1997, pp. 557–560.

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

[15] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Trans. on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[16] C. Clavel, I. Vasilescu, G. Richard, and L. Devillers, "Voiced and unvoiced content of fear-type emotions in the safe corpus," *Proc. of Speech Prosody*, Dresden, 2006.

[17] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Speech ermotion recognition separately from voiced and unvoiced sound for emotional interaction robot," in *International Conference on Control, Automation and Systems*, 2008, pp. 2014–2019.

[18] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Time-scale Modifications based on an Adaptive Harmonic Model," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Vancouver, CA, May 2013.

[19] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Pitch modifications of speech based on an adaptive harmonic model," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2014,.

[20] G. P. Kafentzis, O. Rosec, and Y. Stylianou, "On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models," in *Interspeech*, Portland, Oregon, USA, September 2013.

[21] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Stanford University, 1999.

[22] Y. Agiomyrgiannakis and O. Rosec, "ARX-LF-based source-filter methods for voice modification and transformation," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.

[23] A. Spanias, "Speech Coding: A tutorial review," *Proceeding of the IEEE*, vol. 82, pp. 1541–1582, October 1994.

[24] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An Extension of the Adaptive Quasi-Harmonic Model," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Kyoto, March 2012.

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

[26] M. Kepesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech," *Speech Communication*, vol. 48, pp. 474–492, 2006.

[27] L. Weruaga and M. Kepesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, 2007.

[28] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Munich, Apr 1997, pp. 1303–1306.

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

[30] G. P. Kafentzis, O. Rosec, and Y. Stylianou, "Robust full-band adaptive sinusoidal analysis and synthesis of speech," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2014.

[31] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of Acoustical Society of America (JASA)*, vol. 124, pp. 1628–1652, 2008.