Voice Quality Modelling with the Wide-Band Harmonic Sinusoidal Modelling Algorithm

STYLIANOS IOANNIS MIMILAKIS

MASTER THESIS UPF / 2014

MASTER IN SOUND & MUSIC COMPUTING

Master Thesis Supervisor :
Dr. Jordi Bonada
Department of Information & Communication Technologies
Universitat Pompeu Fabra, Barcelona

He with the rough voice...

... performed in LaTeX

ABSTRACT

Modern advances in the areas of speech and voice processing, have underlined the significance of voice qualities. These qualities, have been proved to provide an increased perceivable naturalness in applications spanning from text to speech synthesis and sound source separation to singing voice conversions and transformations. As a result, different and multiple approaches co-exist, with main task to reproduce and transmit these specific voice characteristics. In this work, we aim to model these voice qualities incorporating robust analysis algorithms, alongside with machine learning tasks. This methodology, allows the extraction and modelling of specific features and patterns, that are enabling the re-synthesis of the phenomena involved during each voice quality. Then, the extracted patterns are fed into an ensemble of Artificial Neural Networks training procedure, capable of generalisation and satisfactory performance among restricted audio corpus. Finally, for the final transformation stage each input voice is activating the Artificial Neural Networks enabling and predicting the re-synthesis of the voice qualities patterns, while allowing the operation to perform in an adaptive way. The proposed method was also evaluated through series of subjective listening tests, where a set of singing voices was processed and 8 experienced listeners had to rate the perceived naturalness, expressivity and transparency of each audio segment. Results are demonstrating the solid performance, achieving almost adequate, to original audio corpus, perceived naturalness, while the perceptual expressivity grade was higher for the transformed audio corpus. As far it concerns the transparency, a mean total success rate of 47.2% was achieved, during the distinction between original natural voices and transformed ones.

Keywords : *Singing Voice Processing & Modelling, Voice Quality(ies), Wide-Band Harmonic Sinusoidal Modelling, Artificial Neural Networks*

ACKNOWLEDGEMENTS

As it is well known, sections like this one, always contain a "laundry-based" list of names that gratitude is being given upon. Alas, boring but bare with me for a moment to accredit persons that really were important to me these last months.

First of all, i would like to thank Pr. Xavier Serra for giving me the opportunity to join the Sound & Music Computing master program, which granted me tons and tones of experiences. Moreover, Dr. Jordi Janer for his helpful comments and observations. And last but not least, Dr. Jordi Bonada for introducing me to the majestic world of voice processing, enduring me and my inquiries and enlightening my path throughout the whole research period. My sincere gratitude.

Finally, without the support of the following persons, failure would have dethroned me.

Friends & Family : Panagiotis, Persephone, Theda, Harris, Lina, Kyriaki, Evi, Carmen, Constantinos.

Colleagues : Konstantinos, Dionysios, Andreas.

SMC 13-14 Creatures : Aram, Oriol, Rogerr, Jorge, Dani.

Wholehearted appreciation of their encouragement..

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ANN(s) :	Artificial Neural Network(s)
APS :	Artefact related Perceptual Score
ASR :	Automatic Speech Recognition
CBR :	Case Base Reasoning system
EES :	Expressive Speech Synthesis
EM :	Expectation Maximisation algorithm
EVE :	Extreme Vocal Effects
FD :	Frequency Domain
FOFS :	Formant Wave Function Synthesis
GCI :	Glottal Closure Instants
GMM :	Gaussian Mixture Modelling
GUI :	Graphical User Interface
HMM :	Hidden Markov Models
HMN :	Harmonic plus Noise Model
HNR :	Harmonic to Noise Ratio
LSF :	Line Spectral Frequencies
MFCC :	Mel-Frequency Cepstral Coefficients
MFPA :	Maximally Flat Phase Alignment
LP :	Linear Predictive
LPC :	Linear Predictive Coding
OLA :	Overlap and Add Method
PCA :	Principal Component Analysis
PSOLA :	Pitch Synchronous Overlap and Add Method
PSTS :	Pitch Synchronous Time Scaling

RMSE	Root-mean squared error
SVLN :	Separation of the Vocal-tract with the Liljencrants-Fant Model plus Noise
TTS :	Text To Speech synthesis
VoQ(s) :	Voice Quality(ies)
WBHSM :	Wide-Band Harmonic Sinusoidal Modelling

1

INTRODUCTION

Speech signals transmit a wide range of information. Apart from utterances that are manifesting messages of principal information, an underlying layer of details is also existent and important in oral communications. This underlying layer, reveals the speaker identity, which is vital for differentiating between multiple speakers [1].

The perceived characteristics incorporated in the aforementioned layer, are derived from a variety of of laryngeal and supra-laryngeal features, which are not unique to one individual speaker, but they are able to form clusters of identifiable voice types [2]. Namely, clusters of *modal, breathy, pressed/tensed, creaky, nasal and harsh/rough* voices.

Assuming the above, it is eminent that the reference to voice qualities is prompting to the definition of the effect, that is being produced by specific vocal tract and laryngeal anatomy alongside with specific vocal routines. Hence, these procedures will form the perceivable voice types (*modal, breathy, etc.*) and then the types will be the conveyor of perceptual information, enabling the individual speaker identification or the transmission of emotions [2].

Besides the emotion transmission and speaker identification, these voice styles are taking place into more artistic areas, such as singing performance. Where in this context, different music genres were observed to include them as main medium of expression [3].

Subsequently, by Voice Quality(ies) or VoQ(s) we will be referred to the perceived voice types, that are formed by specific vocal anatomy routines, in the same way as the cited literature.

In order to approach the examination and analysis of the above phenomena, several methodologies have been proposed. These methodologies, can be classified into three major groups [4] :

1. *Videoendoscopy methods, using image or video data recordings of the vocal folds, such as videokymography.*

2. *Electroglottography (EGG) methods, using measurements of electrical resistance between two electrodes placed around the neck, through which the vocal fold contact area can be estimated.*

3. *Acoustic analysis methods, using audio recordings of the radiated speech signal to compute parameters related to voice.*

By simply comparing the above groups, acoustic analysis seems to overcome drawbacks that the other two methods are exhibiting. More specifically, acoustic analysis methods have lower computational and financial cost and are non-invasive. Moreover, they can generate quantitive results, which can be used for unsupervised tasks [4].

Presuming the advantages of acoustic analyses, the main goal of this thesis, is to examine, model and re-synthesise multiple voice qualities utilising robust analysis / re-synthesis frameworks in conjunction with machine learning modi operandi. More specifically, the proposed methodology avails Wide-Band Harmonic Sinusoidal Modelling [5], allowing the exploitation of voice qualities patterns, which are used for entraining an ensemble of Artificial Neural Networks. Thus, enabling an extensive examination and prediction of the occurred phenomena, in the contexts of parameterised re-synthesis. This will overwhelm the limitations of current approaches, which are exemplar concatenation, scaling and one versus one voice conversion / spectral transformation [6, 7].

The following research work is focusing upon types of rough voices denoted as *growl* and *creaky*, while it achieves the implementation of a robust system capable of learning and predicting specific parameters, that are synthesising phenomena that can be used for transforming an input voice signal, with a greater focus on singing voice performances.

In conclusion, the rest of the document is organised as follows: Chapter 2 provides a synopsis of state of the art technologies in the scope of the current thesis, while Chapter 3 describes the structure of the employed algorithm. In addition to this, Chapter 4 introduces the reader to the proposed methodology of analysis and modelling. Chapter 5, describes the followed experimental procedure, where it's results are being demonstrated in Chapter 6. Finally, Chapter 7 concludes this document by denoting the future work.

2

RESEARCH & TECHNOLOGY SYNOPSIS

2.1 PREFACE

In this chapter, the main research areas, that this thesis is covering, are being demonstrated. More specifically, state of the art approaches and relevant technologies, considering analyses, emulation and re-synthesis models, are investigated throughout the following sections.

The first section entitled "Physiological Mechanism", takes place into the investigation sphere of physical phenomena , regarding human utterances and how the establishment of the voice qualities term, took place. Next, "Analysis/Synthesis Model" is concerned with the discipline of archetype, inside digital signal processing, for the analysis and re-synthesis of voice signals, allowing a range of applications, spanning from spectral emulations to transformations and conversions.

Assuming that the Analysis/Synthesis models, have provided feasible solutions to the emulation of the aforementioned phenomena, another area of research is being introduced. More precisely, this field is being denoted as "Emulation" and involves the overview of approaches that try resemble parallelism between different physiological mechanisms. An illustration of the co-existence and coherence of these different areas, can be previewed in Figure 1.

Finally, industry has also advanced the engines, regarding the processes of voice analysis, re-synthesis and transformations. As a result, they can not be neglected from this study and the last section is dedicated to the exploration of the capabilities of these machineries.

2.2 PHYSIOLOGICAL MECHANISM

Voice quality variations, as a set of voice modifications, have been studied from the mid 1960's. Thus, a great gamma of perspectives has been established.

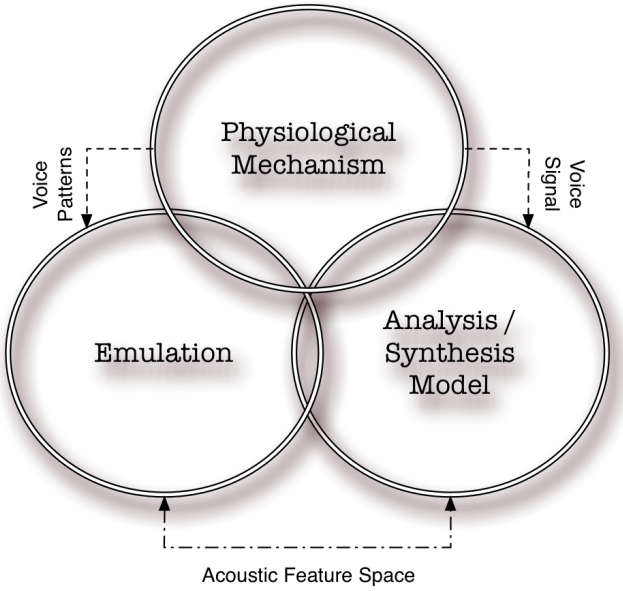


Figure 1: Examined Research Areas & Relationship.

More specifically, the first one that brought to light these quality varieties, was D. Abercrombie in [8], where the alteration in simple phonetics alongside different utterances and speakers, was examined. The above research, triggered the interest of relative fields and as a result in [9], the previous variations were examined from the physics of human's larynx system. This examination took place alongside cross-language subjects. As an outcome, it was the discrimination of larynx usage in three distinctive procedures.

Focusing on the actual procedures, the first one affects the fundamental frequency of voicing, by modifying the laryngeal tension. The second, is concerned with the timing onset of supra-glottal movements, during voicing, for the determination of pre-voiced and voiceless, (un)/aspirated consonants. Finally, the third one, which is the separation between arytenoid cartilages, classifies the utterances into voicing modes such as breathy, creaky, modal, voiceless, etc [9].

According to [10], the previous classification of voicing modes triggered the research of these physiological phenomena. These studies spanned from laryngoscopical examinations up to analyses of acoustical parameters and properties. The most significant one, from the part of laryngoscopical examinations, was done by J.Laver in [11].

Analytically speaking, in [11] the physiological correlation with the above voicing modes, henceforth described as voice qualities (VoQ), were studied in terms of muscular tension. Therefore, three distinct parameters of tension were

devised, that can describe each VoQ. Namely, *adductive tension*(the action of the inner arytenoid muscles adducting arytenoids), *medial compression* (the force on the vocal processes adducting the glottis) and *longitudinal tension*(the tension of the vocal folds themselves).

The proposed parameters atop, offered an analytical explanation of each individual VoQ. Technically speaking, *breathy voice* involves minimal laryngeal tension, while vocal fold vibration is inefficient and the disunion of vocal folds is resulting an audible frication noise. Following a similar pattern of excitation, *whispery voice* is characterised by low tension of the inner arytenoid muscles combined with a high medial compression, while laryngeal vibration is once again very inefficient.

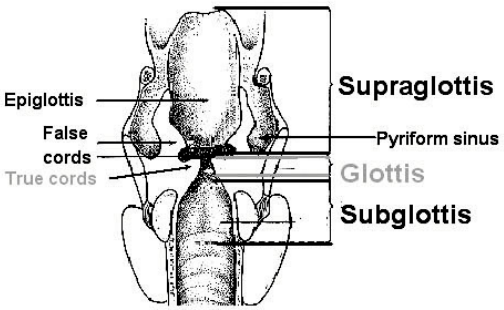
Moreover, a *modal voice* can be described as a moderate laryngeal tension, among with efficient vocal fold vibration, while the cartilaginous parts of the glottis are vibrating as a separate single unit. *Tense voice*, has a higher degree of tension alongside the vocal tract with the implication of adductive tension and medial compression [11].

On the other hand, *harsh and creaky voices* involve high tension settings. With defining characteristics be endowed with additional a-periodicities, due to these high glottal tensions. In addition to this, high medial compression and adductive, but low longitudinal tension, provide two more distinctive features [11].

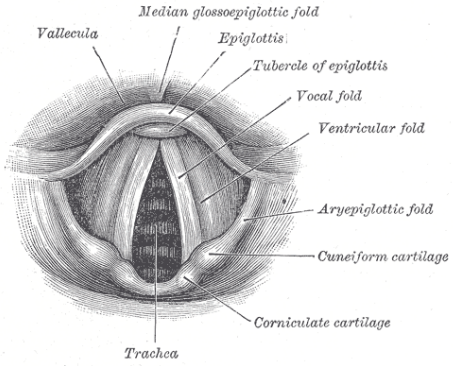
An illustration of the different parts involved in human's voicing mechanism, denoted alongside the document, is being given in the following Figure 2

Focusing on the latter case of acoustical properties analyses, a key research is being introduced in [10]. In this article, different voice qualities are complementary examined among female and male speakers. Main results were differences in potential acoustic cues, between VoQs. These differences included a high increase to the following features of:

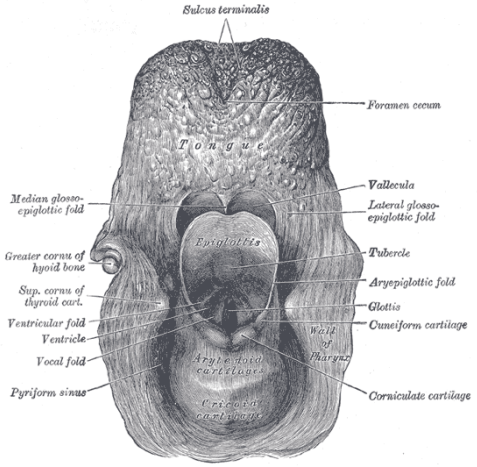
1. Relative amplitude of fundamental frequency component, as a proportion of the period of open glottis is being increased.
2. The amount of aspiration noise, that replaces higher frequency harmonics as the arytenoids become more separated.
3. Lower formant bandwidths.
4. Complexity of the frequency response, in terms of deviations, which is associated with the tracheal coupling.



(a) Overview of the Larynx



(b) Larynx Front View Entrance



(c) Larynx Back View Entrance

Figure 2: Human Larynx.

Additional observations were, that on average, female voices tend to be more breathy than male ones. Additionally, a great number of utterances are ending with a breathy type of vibration among with diplo-phonic irregularities in the timing of glottal periods.

Finally, these acoustic cues were synthesised, using a common inverse filter approach, for perceptual assessment. This subjective evaluation, showed the importance of these differences, underlining the significance of diplo-phonia and other related deviations of perfect periodicity, for achieving more natural sounding synthesis.

Beside the stated voice qualities, laryngoscopic observations in [12], showed that a renewed definition, regarding places and nature of articulation, was emerging. In essence, that pharyngeal manners of articulation go beyond fricative approximations and it should be handled as a parallel or accompanying phonological process. Meaning that, from now on a facilitation of a more precise

explanation of phonological system and more detailed description of phonemes, could be established. This hypothesis was validated through subjective listening tests, using synthesised audio corpus. More precisely, the subjects denoted a higher preference rate, for the synthetic audio that was taking into account pharyngeal features.

These new advances in articulation system, expanded the gamma of different perspectives. As a result, collateral research areas begun to gain interest. As an example, in [3], a specific singing voice technique, which was observed alongside different music genres, was examined. This technique, was denoted as *growl* and the authors in [3], claim that it can easily be perceptually related with the *creaky, harsh,* or *rough* voice.

A throughout investigation in [3], regarding this VoQ, indicated growl as the main aspect of ethnic music spanning from South Africa to Japan and Mongolia[3, 13], as well in different music genres such as jazz, pop and samba.

Moreover, analyses of the actual phenomenon of growl, using video - fluoroscopy and high-speed imaging among with the mapping of it's acoustical characteristics, are also inquired by the same authors in [3]. The performed analyses stages encountered high vibration and position of the larynx and aryepiglottic folds. As a result of aryepiglottic folds vibration, the vocal tract is being given a unique shape, constricting the larynx tube and then the growl is produced. As far it concerns the examination of acoustical features of growl, a predomination of sub-harmonic generation and high oscillation patters was observed throughout all different singing voice techniques. Meaning that, relation awareness is crucial between perceptual cognition in different traditions.

These perceptual relations between voice qualities, studied in the literature covered in this sector, have also provided "fertile soil" for another scope of analyses and estimations. This scope can be discriminated itself into two categories. The first one including probabilistic estimations for the objective determination of voice qualities [14, 15], while the other focuses on the cognitive aspects of emotion and mood [16]. Meaning that having such a series of overlapping research fields, the emulation of the voice qualities is implicated more and more into the level of significance and importance. Thus, the following part of Chapter 2 is concerned with the synopsis of methods involved into emulation of these qualities.

2.3 EMULATION

In the previous section, the physiological mechanism was the main subject under examination. Most of the described approaches above, used methods of artificial synthesis for each corresponding VoQ. Therefore, these synthesis tasks must not be confused with the emulation of phenomenon, since the main task at this point, is to obtain patterns from different aspects of analysis, that can be applied to a target voice signal.

In addition to this, the importance of these tasks can be overviewed by the product of this fieldwork, allowing the unfolding of new scopes of experimentation. Which consequently, provided a diverse range of applications, spanning from enhanced emotional speech synthesis [17] up to improved source separation[18].

Onwards this notion, the first attempt of emulation of such phenomena is introduced in [19]. The current approach uses a combination of two physical models, one for vocal fold and another one for vocal tract modelling. The main reason of the examination of the fusion of these two models, is the rigorously characterisation and description of each VoQ.

As a matter of fact, area functions, provided from different physiological examinations, were used to derive a statistical correlation, of observations, using Principal Component Analysis (PCA). This analysis showed the cross-sectional activations of different laryngeal parts up to the lips and patterns of deformations in the vocal tract. Finally, these activations were mapped, according standard deviations of each component, and tested for their interconnection with different audio-based recorded VoQs.

This operation, resulted into simple spectral characteristics, such as fundamental frequency deviations. During the same utterances a pointed drawback, was that these models fail to fully reproduce the naturalness of original recorded VoQs. Still, the enhancement of speech recognition systems was achieved [19]. For more details, regarding both models, the reader is encouraged to inquire into the previous section and the related article in [19].

Aside from physical modelling methods, prosperous tactics have also been provided by digital signal processing oriented ones. These methods have overcome limitations of former ones and clearly, they are setting higher goals. A simple explanation to this escalation, is the significant contribution of advanced analysis and synthesis methods, that are fully described in the forthcoming section.

Focusing on the actual emulation strategies, a system that models the characteristics of roughness and growling, in voiced signals, is being presented in [20]. More specifically, the unavailability of real time voice transformations of time domain (TD) methods, is outlined and a direct approach in the frequency domain (FD) is being proposed. As a consequence, a procedure of pure sinusoidal addition inside spectrum, to sub-harmonic frequency regions and phase peak shifting with specific offset, reproduces desired down octave transpositions. With the main intention that, these transpositions can assemble the perceptual discrimination of a rough voice.

At this stage it should be stated, that only sub-harmonics inside the range of estimated fundamental frequency up to 8 kHz bands are admixed to the spectrum. An explanation to this, is that higher sub-harmonics are irrelevant, in terms of perceptual reproduction of the roughness [20]. In addition to this, observations showed that the magnitude of fundamental frequency masked the first sub-harmonic.

According to [3], sub-harmonic generation could be easily related with another VoQ, denoted as growl. Consequently, a similar approach is also used for the growl emulation, in [20]. The main differences between rough and growl emulation, are the specific amount of sub-harmonic addition and different peak phase alignment, due to the existence of individual macro periods, in the case of growling. For the implementation above, a phase-locked vocoder technique was used.

The above approach, was revisited in [7]. This time, an exemplar based procedure takes place, in order to replace the sub-harmonic generation process. Consequently, a typical phase vocoded implementation was used for spectral analysis and mapping of the target, modal and source, growl voice signals.

This not only allowed robust growling emulation and re-synthesis, but also a blistery procedure in situations where no great audio corpus is available [7]. Finally, the approach was assessed by listening tests, using natural and re-synthesised growls, in terms of ‘naturalness”, quality and voice expressiveness. Results, showed a great ambiguity, of experienced listeners, in distinguishing between real and synthetic growls and an overall good performance alongside the constrains of expressiveness and quality.

Inspecting VoQ emulation by a perceptual aspect, an approach of breathy vowel emulation is proposed in [21]. This method, operates in the FD and tries to decompose the signal into a periodic and a noise component. This will allow the modification of the envelope of the noise component, which is a function

of the glottal flow waveform and thus, the assessment of different VoQs. The algorithm for the decomposition used was the one described in [22].

Focusing on the actual study case, breathy vowel was simulated with Liljencrants-Fant model (SVLN) [23] and analysed using the above decomposition method. Preliminary analysis showed that breathy noise source can be characterised by strong modulations based on a "gating" function, resulted by the vocal fold oscillations. Hence, four different patterns of oscillation were scoped out and assessed by subjective listening tests. At that point synthetic and recorded vowels were served as input to the proposed system and the subjects had to rate the output, in terms of perceived naturalness.

The assessment results, showed that a DC based oscillation, which is similar to physiological glottal airflow, was preferred alongside the subjects. Another interesting fact, was the statement of complexity of breathiness phenomenon and it's emulation/modification procedure. Where, in the search of superior emulation process, more precise and robust analysis / synthesis models should be taken into account [21].

Onwards this idea, a more robust model was presented in [24], where a FD approach combined with pre-processing stages, labeled as adaptive pre-emphasis, enhances the LP operation, yielding better transformed results. The output of the enhanced LP model, was validated once again with subjective listening experiments, by achieving a well balanced identification rate only for breathiness emulation and transformation.

Remaining in the scope of perceptual aspects of VoQs, emulations in context of expressiveness projected upon emotional states, were investigated in [17]. The main goals of that research were two. First, to demonstrate the significant improvement of VoQ emulations combined with prosody modifications, for transforming a neutral to an expressive speech style, in contrary to previous methods using only prosody adjustments. And the proposal of a methodology capable of measuring parameters related to different VoQs.

The measuring methodology, yielded a set of useful metrics. More specifically, the following list describes and gives an overview of these metrics.

1. *Jitter & Shimmer* : Denoting cycle-to-cycle variations of the fundamental period and waveform amplitude, describing frequency and amplitude modulation characteristics.

2. *Harmonic-to-Noise Ratio (HNR)* : Describing the ratio between harmonic and stochastic components.

3. *Hammarberg Index* : Defining the difference between the maximum energy in the $0-2$kHz and $2-5$kHz frequency bands.

4. *Relative amount of energy* : Computing the ratio of energy for frequencies below and above 1kHz of a voice signal spectrum.

5. *Spectral Energy Drop-off* : Reflecting a spectral "tilt" above 1kHz.

At this stage, it should be mentioned that in [27], these metrics were redefined. Ás a result, mean and standard deviation values were computed for each recorded audio sample, containing a single VoQ .

As far it concerns the implementation details, a three stage approach takes place into action. The first stage, uses a harmonic plus noise (HMN) decomposition based on [25],that operates in the FD. Unfortunately, the previous implementation does not preserve "harmonicity" of the spectral components [17] and for that reason an optimisation procedure, described in [26], is introduced implicating the overall computational procedure.

The second stage, is concerned with the emulation of VoQ and the feature extraction for the prosody prediction. To do so, a data mining system, entitled Case Base Reasoning (CBR), obtains prior and target information for each phoneme and related VoQ parameters are selected. Finally, the speech transformation is carried out, based on the results of the HNM analysis / synthesis and the selected parameters are concatenated on top of each harmonic's magnitude, phase and energy contour information.

To this end, the proposed system in [17] was evaluated, in terms of subjective listening tests, according to preference rate of perceptual emotional parsing between prosody modifications and the combination of VoQ modelling. The audio corpus used for the emotional VoQ retrieval, was the one presented in [28], including neutral, happy, sad, etc. recorded samples representing different correspondent VoQs.

Results, showed that the combination of VoQ characteristics and prosody parameters, produces a more preferable, by the listeners, stimuli. Moreover, the additional emulations of VoQ, improve the perception of all expressive and speech styles.

On the same tracks of voice expressiveness, a specific focus on singing voice in extreme music genres, was given in [29]. This approach revealed patterns and characteristics of physiological aspects, among with a feasible strategy of emulation, for the transformation of neutral voice signals.

In more detail, user assisted recordings were conducted and each type of singing voice, that contained a specific VoQ, fell into one of these categories: *Distortion, Rattle, Growl, Grunt, Scream.*

Then, each individual recording was fed to a robust FD model of analysis and synthesis, that not only preserves the "harmonicity", but also allows a great variety of voice transformations in real time [5]. Afterwards, spectral envelopes and macro-periodical patterns were collected and used for transforming neutral voices.

Most of the above researches, have outlined two major contributing factors. The first one is concerned with the undertaken analysis / synthesis model, affecting the output signal's overall quality, while the second one pinpoints the initial strategy of acquiring and parsing parameters of each VoQ, with main effect the perceived naturalness.

Alongside this belief, different approach incorporating voice conversion techniques, is introduced in [30]. Technically speaking, spectral and prosodical parameters were extracted using a FD method [31] and different voice conversion algorithms were studied. Namely, weighted codebook mapping, an extension of it denoted as weighted frame mapping and a Gaussian Mixture Modelling (GMM) framework, were the three algorithms under investigation.

The three aforementioned algorithms, extracted the information for each specific VoQ from the database in [32] and transformed neutral parallel utterances, using the extracted prosodical / spectral information. Finally, a series of objective and subjective evaluation took place. The objective evaluation, was based on the computation of root-mean squared error (RMSE) of Bark-scaled line spectral frequencies (LSF), between the neutral and converted voice signals. As for the subjective assessment, listening experiments were held, at which' the listeners had to identify the perceptual mood(aggressive, cheerful, depressed).

Conclusions of the experimental procedure, whereas GMM and weighted frame mapping frameworks showed a better response, in terms of RMSE and high recognition of emotional speech.

2.4 ANALYSIS / SYNTHESIS MODEL

Throughout the previous section, all of the emulation procedures relied upon a model, that allowed the analysis, the acquisition of spectral and prosody information and the final re-synthesis of an utterance. This differentiates with voice source models, which their task is to synthesise voice signals. For that reason, these two models should be treated separately and not be confused in conjunction.

As a consequence, the main goal of this section is to provide an overview of the most relevant models available, escorted by their foundational advantages and drawbacks. By inspecting the inner mechanism of these models, one could easily discriminate them into two major categories. The first one is characterised for it's operation in the time domain (TD) and the second one for taking place inside the frequency domain (FD).

In addition to this, each one of the below approaches, either in TD or FD, can also be discriminated as a narrow or wide band procedure, depending on the temporal resolution they are taking into account for the analysis part. Meaning that if the analysis window is covering one or two periods of the signal, the wide-band conditions are met and vice versa. The following Table 1, illustrates the procedural condition of each model.

The following subsections, are focusing on the description of each model of the above Table 1, according to the two class discrimination.

2.4.1 *Time Domain Mechanisms*

By reviewing the TD class of methods, the most widely known is the one described in [33, 34]. More specifically, it is a variation of a time stretching algorithm, denoted as PSOLA, that specialises into monophonic sounds that can be characterised by pitch, such as the case of human voice.

This algorithm, first exploits knowledge of the fundamental frequency and pitch, in order to correctly synchronise time segments, avoiding any discontinuities in pitch. Next, a time-stretched version is being synthesised by overlapping and adding the synchronised time segments at different time instants [34].

The main drawback of a highly pitch-dependent method, such as the one described above, is the existence of non-periodic components, like consonants, which can produce artefacts.

MODEL ACRON.	NARROW-BAND CONDITION	WIDE-BAND CONDITION
HQSM	✗	✓
PSOLA	✗	✓
FD-PSOLA	✗	✓
LP-PSOLA	✗	✓
PSHF	✗	✓
Phase Vocoder, Sinusoidal Modelling & Spectral Envelope Implementations	✓	✗
SHIP	✓	✗
STRAIGHT	✗	✓
WBHSM	✗	✓

Table 1: Band conditions satisfaction

Onwards improving the performance of the previous PSOLA algorithm, another method in [35], denoted as LP-PSOLA is being presented. This time, the input voice signal is being filtered, in order to obtain formant shapes. Then the voice is being parsed to the typical PSOLA model and the filtered formants are summed back again, after the overlapping addition.

Unfortunately, quality degradations were observed of when greater fundamental frequency modifications were applied. The observation was originally made by the authors of [36] and an improved model is proposed. This advance, was based on the encapsulation of LP residual decomposition for each period. Meaning that, the successful simulation of most relevant glottal source parameters could be acquired. In addition to this, more precise modifications to the periodic and non-periodic components could be established[36].

One main disadvantage of all the atop TD models, is that they do not allow the modification of each individual glottal pulse [5]. A method that allows this operation, is the one described in [37], which incorporates a flexible and fast additive synthesis engine based on Formant Wave Function Synthesis (FOFS).

The method above models the human voice as a combination of impulses, equivalent to the vocal chords and a set of band-pass filters, representing the characteristics of the vocal tract. However, this approach does not precisely

represent the amplitude spectrum, neither phase alignment is being taken into account [5, 38].

A crucial factor of all the above TD models, is that they do not allow a great range of timbral transformations [5]. As a result, FD approaches are emerging towards the necessity of diversity [39].

2.4.2 *Frequency Domain Mechanisms*

The successiveness of PSOLA methods encouraged the advances of similar approaches. As a result, the method in [40](FD-PSOLA), performs almost the same operation, but in advance, a series of Short Time Fourier Transform (STFT) analysis takes place atop. Still, the modification of the spectral components becomes cumbersome, where no peak indication is existent.

Aiming for peak location retrieval to simplify the process, another model is being presented in [39, 41]. With respect to the technical aspects, this approach performs a STFT analysis and determines the fundamental frequency by cross-correlating a compressed version of the magnitude, with a series of comb-filters, corresponding to various pitch candidates. Then, peak selection procedure takes place in, where given a fundamental frequency at each frame, multiples of fundamental frequency are computed, so that spectral envelopes are easily derived.

Using the above method to reconstruct voice signals, yield some undesirable effects [39], often described as "phasiness". This undesirable artefact usually affects most frequency-domain techniques and has been linked to the lack of phase synchronisation [39].

Endeavour onwards better re-synthesis quality, an approach that combines the advantages of TD and FD models into a single framework, is introduced in [42]. This method, provides an independent control of each glottal pulse and flexible phase/timbre modifications. In addition to this, phase alignment conditions are also met, reducing any artefacts of the re-synthesis procedure.

A latter advance of the above algorithm, is proposed in [5], denoted as WBHSM. With this approach, a single or double period of the analysed signal is used to estimate individual harmonic components, allowing higher temporal resolution with various timbral adjustments, performed in real-time. Another interesting point is that periodic or even non-periodic components, of voiced utterances, are solely represented by sinusoids [5].

Similar to the WBHSM model, another approach is presented in [43] indicated as (HQSM), that tends to enhance the performance of another system entitled STRAIGHT [44]. As a result, a simpler, more robust and less costly, in terms of computational resources, model is implemented. The crucial of that implementation, is the usage of dithering noise to enhance the output re-synthesis perceived quality, allowing a great range of spectral manipulations, even in cases of extreme modifications [43].

Main drawbacks of the aforementioned implementation, are the additive noise affecting the amplitude and phase of higher harmonics, thus smearing the yielded envelope, the denial of control of each individual glottal pulse and the unavailability of a robust real-time implementation [45].

Emphasising the importance of low computational models, an approach that does not need any information about the fundamental frequency or onsets of individual glottal pulses is presented in [45]. The algorithm distinguished as SHIP, uses a phase vocoder implementation and tends to obtain optimal time shifts, by cross-correlating past input time frames. As a result, it preserves phase alignment with minimal computational costs. Apart from the phase vocoder approach for the alignment, the rest of structure is similar with typical SOLA methods [45].

Finally, limitations of the above model can be observed during the assessment in [45]. At which' the SHIP algorithm can not grasp with pitch transpositions of male voices that exceed the factor of 2, which is a typical procedure of basic gender alteration (male to female).

2.5 INDUSTRIAL ENGINES

The first chronologically stand-alone application that performed vocal synthesis, based on physiological models and wave guide based vocal tract mode, is described in [46] (SPASM). In addition to the vocal synthesis, it also allowed adjustments to produce different VoQs similar to roughness and growling.

Another commercial application, produced by Antares is THROAT [47]. This implementation, allows the emulation of roughness and growl voices, by simple modulations Moreover, breathiness is also emulated by controlling variable frequency noise shaping that is mixed back to input signal.

One of the most accurate emulations of growling is the one that TC-Helicon, under the product alias VoicePro [48]. Still, it does not create extreme growl or rough emulations [29].

Finally, a collaboration with Music Technology Group and Yamaha, yielded a FD approach, named KaleiVoiceCope, for not only transforming voices in real-time, but also emulating rough and growl voices. based on the approach in [20].

AFTERTHOUGHT

Assuming all the above models, mechanisms and methodologies, presented alongside Chapter 2, in conjunction with the outcome of [29], regarding the qualities of commercial applications, it can been seen that the necessity of a multi prism research, is emerging. Thereby, the following research produced in this M.Sc. thesis assumes the following factors for succeeding the joint and overall modelling of voice qualities:

1. Robust Analysis / Synthesis Model, preserving a balance between the condition satisfaction and computational efficiency, for real time applications.
2. Sanction of the diversity of transformation strategies.
3. Better understanding of physiological behaviours, providing preferable emulation strategies.
4. Robust utilisation of the emulation strategies, for achieving precise models, presuming the advances of machine learning techniques.
5. Quality assured applications, with respect to the music recording, mixing, and producing community.

3

WIDE-BAND HARMONIC SINUSOIDAL MODELLING

3.1 PROLOGUE

In the previous Chapter 2, a synopsis of the state of the art technologies and current needs were presented. One of the denoted aspects, was the incorporation of a robust analysis and synthesis model, preserving homogeneity between computational efficiency and satisfactory inner processing procedure.

For this reason, current work incorporates the WBHSM and this Chapter 3 is concerned with the description of the algorithm involved. In addition to this, sub-systems allowing the extensive analysis and re-synthesis strategy are also taken into account. As a result, the following Sections are concerned with three major operations, involved during analysis and re-synthesis of an input voice signal.

More specifically, the main structure of the WBHSM algorithm consists three main steps :

1. Compute Fundamental Frequency
2. Perform Maximally Flat Phase Alignment (MFPA)
3. Perform Sinusoidal Modelling in Wide-Band Conditions

For each one of the above steps, a specific section exists for further implementation details.

3.2 FUNDAMENTAL FREQUENCY ESTIMATION

The procedure of fundamental frequency estimation has often been in the centre of attention and a lot of approaches have been presented in the last years. Most of the proposed methodologies underline the significance of acquiring robust estimations, in a target range of applications spanning from audio effects to music analysis and transcription [50, 51].

Throughout the existing research, a prominent issue of the frequency fluctuations that even can hop from one musical octave to another one [50] is discussed. For solving this inherited problem, tasks of validation and processing stages have been encapsulated during the procedure of computation. One important factor is that these tasks are prone to alter depending on the input signal [50, 51].

Since this current work is concerned with human voice signals, the fundamental frequency estimation is based on a spectral autocorrelation method described in [51]. More specifically, the input audio signal is being down sampled up the rate of around 11kHz, in order to decrease the computational time. The technique for downsampling is based on a set of poly-phase filters allowing a precise sample decimation. As a next step, the down sampled signal is being sliced up to consequent smaller proportions, of length $\simeq$ 50 ms at a rate of $\simeq$ 172 frames per second. Then, each portion is being transformed to the frequency domain using a Blackman-Harris window, without any intermediate processing such as zero-padding.

The exported transformation containing the complex - valued spectrum is being filtered, in order to provide a multi-resolution spectrum. The filtering procedure consists a convolution of each frame with a set of triangular kernels of variable length, likewise 0Hz length at 0Hz and 72Hz length kernes at bins of 1 kHz and above. Afterwards, from the multi-resolution spectrum the logarithmic amplitude values are computed and a smoothing average procedure takes place. This procedure incorporates variable triangular windows, where at that moment the windows are having 80Hz length at 0Hz up to 180Hz length at the frequencies of 700Hz and above.

The above smoothing, is followed by a difference function among the logarithmic amplitude spectrum and its smoothed version. At that point, the autocorrelation function takes place and then it is normalised by its maximum value. Finally, local maxima, derived from the previous autocorrelation function, are gathered and fundamental frequency candidates are estimated based on a set of rules, like minimum value, significant autocorrelation value and between a desired frequency range. Where, any great frequency fluctuations and discrimination of voiced and unvoiced parts, are derived from a 2nd order polynomial and audio content features, such as TD zero-crossing rate, respectively.

3.3 THE MAXIMALLY FLAT PHASE ALIGNMENT ALGORITHM

Subsequent to the described step, one would may consider the necessity of fundamental frequency information. Consequently, it has to be stated that by examining a simplified case of voice production, a set of glottal pulses excites the vocal tract, at the specific frequency. This means that each human voice utterances, changes the pitch (i.e. the perceptual fundamental frequency), by modifying the rate which these excitations occur [38].

An observation that gained the attention of many models allowing the transformation of voice signals, was the shape invariance characteristic. This observation was related to the shape of time-domain waveform signal around each glottal pulse onset, that tends to be independent of the fundamental frequency, but it is on the impulsive response of the vocal tract. This means that the transformation in frequency domain, can describe the aforementioned shape in terms of amplitude, frequency and phase values [38].

Assuming the above, it is undemanding to to comprehend the need of matching between detected onsets and actual glottal pulse onsets, in order to obtain the best re-synthesis possible. Thus, different practices for detecting glottal onsets have been developed. As a matter of fact, these methods are relying on setting onsets to arbitrary locations or phase characteristics of the source signal. An approach to estimate glottal pulse onsets, which is undertaken in this work, is being presented in [38].

More specifically, in a constant frame-rate spectral analysis framework, the harmonic phases that fit efficiently are collected when, the property of flat phase envelope under shifts of each formant in a properly centred window, is met [38, 42]. Meaning that, when the analysis window is almost cantered to an actual onset, the harmonics are synchronised and the phase spectrum is nearly flat, with remarkable phase shifts for each formant.

Assuming that whenever a slide shift of a window happens, a phase shift that varies linearly along frequency occurs [38], one path to locate the pulse onset is to estimate the slope of the aforementioned shift. Hence the described algorithm is trying to obtain a maximally flat phase alignment, by minimising the phase differences between harmonics. In practical implementations, phase unwrapping complicates the estimation, so the following steps should be taken into account :

1. Set several arbitrary fundamental phase candidates $\widetilde{\varphi}_0$, in the interval of $[\pi, \pi)$.

2. For each candidate, apply the corresponding phase shift along harmonic peaks, derived from the known linear slide shift of the window.

3. Locate the phase of each harmonic $\varphi_{0,h}$, which will be rotated after the previous step as $\widetilde{\varphi}_{0,h} = \varphi_{0,h} + 2\pi f_h \widetilde{\Delta_t}$, where $\widetilde{\Delta_t}$ denotes the time shift and f_h is the constant frequency of the h^{th} harmonic.

4. Then, compute the sum of rotated phase differences as $\widetilde{\varphi}_{diff} = \sum |princarg(\widetilde{\varphi}_{0,h+1} - \widetilde{\varphi}_{0,h})|$.

5. After computing the sum of differential rotation for each phase candidate, a function is obtained that is similar to a sinusoid. From this function, the minimum value sets the desired fundamental phase φ_{min}, that approximately centres the glottal pulse onset [38].

6. Finally, the closest pulse onset t_{MFPA}, which satisfies the maximally flat phase alignment, can be estimated according to the central time frame t_{frame} by : $t_{MFPA} = t_{frame} + \frac{princarg(\varphi_{min} - \varphi_{0,0})}{2\pi f_0}$.

3.4 THE HARMONIC SINUSOIDAL MODELLING ALGORITHM

In the previous sections, prior steps to the Harmonic Sinusoidal Model were previewed. The main theme of the current section is concerned around the estimation of harmonic components out of a single period of the input signal. This will allow estimating harmonic parameters, with higher temporal resolution than typical phase-vocoder and sinusoidal model based methods [5].

In more details, the current algorithm takes one of the detected period and a windowing function centred to the voice pulse is being applied. The transformation to the frequency domain is being performed using the Fast Fourier Transform (FFT) implementation of Discrete Fourier Transform (DFT), where each bin is corresponding to one harmonic and their magnitude and phase information can be retrieved from :

$$M_k = |\overline{X}_k| \tag{1}$$

$$\Theta_k = \angle \overline{X}_k \tag{2}$$

where, $k = 1, ..., \frac{T}{2}$ is the vector containing the harmonic series and $\overline{X}_k$ is the complex valued spectrum of an input signal $x(n)$, with an estimated period of T and multiplied by a windowing function. The transformation is based on Eq. 3

$$\overline{X}_k = \sum_{k=0}^{N-1} x(n) e^{-j2\pi \frac{kn}{N}} \tag{3}$$

The above equation 3 is equal to $\overline{X}(\frac{kf_s}{N})$, meaning that each bin corresponds to one harmonic, when the following condition is preserved:

$$W(\frac{kf_s}{T}) = 0, \forall k \in [1, T-1] \tag{4}$$

where, $W(\frac{kf_s}{T})$ is the windowing function and the above statement is valid for $N = gT, g \in \mathbb{N}$.

This means that when a window is a multiple of the signal's period, the inter-harmonic energy contribution goes to zero [5]. With the main ambition to achieve widest-band possible conditions, the $N = T$ is used.

In real-world scenarios, the estimated period T may include non integer or power of two values. As a result, the FFT algorithm will require zero-padding, in order to match power of sample lengths, but this will modify the complex spectrum bins in a way that will not correspond to any harmonic components [5].

For this problem, two main solutions have been proposed in [5]. The first one includes a processing stage prior to FFT computation, denoted as periodisation, which attempts to fill the gap of the FFT buffer, by concatenating the selected segment by a specific amount of times derived from period T and then windowed by a proper function. On the other hand, the second approach which this work incorporate, is based on an up-sampling method, where the number of samples matches the closest FFT buffer size M using the above Eq. 5

$$M = 2^{(\lfloor \log_2(T) \rfloor + 1)} \tag{5}$$

and only frequency bins up to the value of $\frac{T}{2}$, are relevant.

So far, only the following procedures have been discussed:

1. Fundamental Frequency Computation
2. Period and Pulse Onset Detection
3. Frequency Domain Analysis in Wide-Band Conditions

What is still missing, are the parts of input signal re-synthesis and the underlining feasible strategies, regarding signal transformations using the described model. As far it concerns the re-synthesis, It is possible to use an equivalent to analysis method, where for each frame the Inverse Fast Fourier Transform (IFFT) is being computed. Extra care should be given at the recovered TD signal, where it has to be down-sampled to the analysis frame rate and overlapped with the other re-synthesised periods, to the exact onset sequence. In addition to this, the phase is being recovered using a minimum phase filter, derived from cepstrum, along with stochastic deviations between similar consecutive voice pulses.

As far it concerns the transformations, we can discriminate them into two major types. The first one related to the period onset sequence and a latter one related to each individual period [5]. Assuming the statements in Chapter 2, this can be related to the typical source-filter voice approach of modelling human voice, with the former class of transformations associated with voice source and the latter with the vocal tract.

This allows combining the operations of TD combined with the diversity of procedures performed in the FD. In addition to this, controlling each pulse onset is enabling a more extensive transformation operation, by simply adjusting or equalising amplitude and phase values. This means that several voice disorders and voice types, that are characterised in the excitation glottal pulse sequence, can be analysed and re-synthesised along time [42].

4

ANALYSIS & MODELLING

4.1 PROLOGUE

In the previous Chapter 3 the structure of the analysis and re-synthesis was inspected. The aforementioned flexibilities that this model allows us, regarding the examination of different voice qualities, will be employed with main ambition to create generalised voice qualities models that are allowing the emulation and re-synthesis of each voice quality.

In the following Figure 3, the proposed methodology for analysing, modelling and transforming voice is being demonstrated.

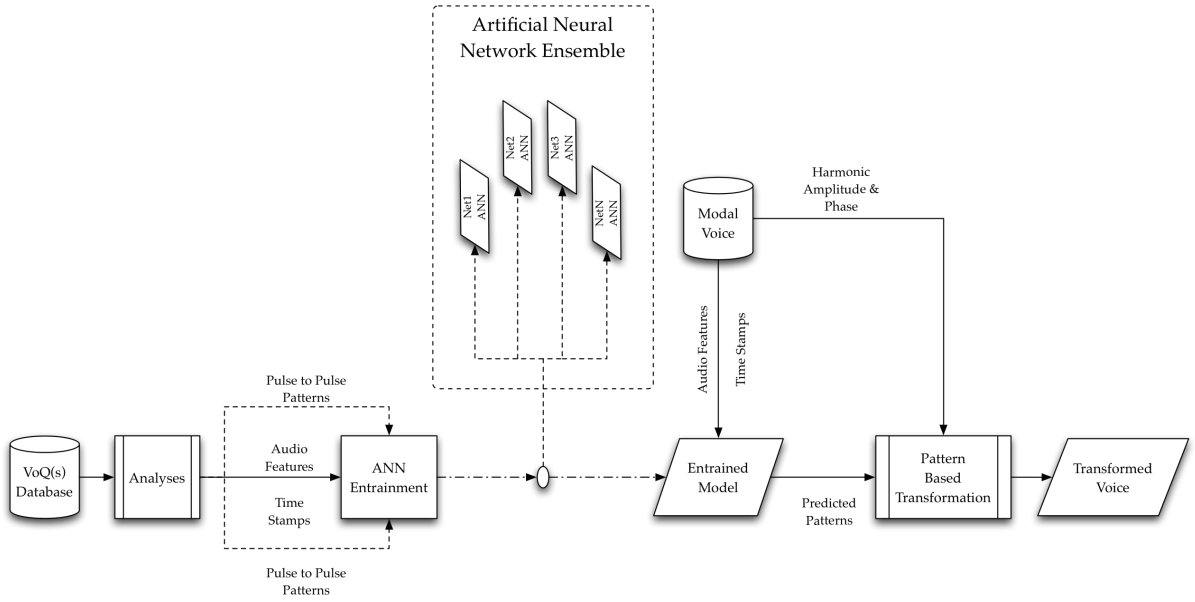


Figure 3: Proposed Modelling Architecture.

Audio recordings containing different VoQ(s) are imported to the framework of WBHSM. The extracted values for each pulse onset, henceforth called voice pulse, are sent to the pre-defined module “Analyses”, which tends to extract patterns that are describing each VoQ.

Then, specific features and time marks, that denote time stamps of of each occurred phenomena inside a VoQ, are used as input variables to a set of multiple Artificial Neural Networks (ANNs), denoted as ANN Ensemble. As far it concerns the output vectors, that are used as target values for prediction, the extracted parameters from the "Analyses" are being set as target. Finally, for each modal voice that is being fed as input to the system, the procedural chain goes as follows:

1. WBHSM Analysis
2. Audio Feature Extraction
3. Feature Prediction
4. VoQ Synthesis, based on the atop predicted features
5. Transformation of each Voice Pulse, according to synthesised VoQ
6. WBHSM Synthesis

yielding the desired transformed modal voice containing the analogous VoQ.

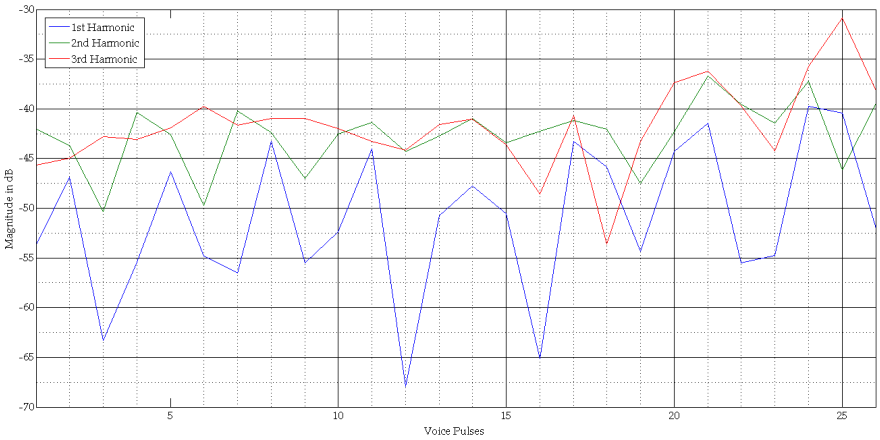
From a general perspective, the whole procedure can be categorised into three main three courses of action. Namely, *Analysis, Training & Synthesis Formulation.*

At this point, it has to be mentioned that for every VoQ, the inner procedures vary for each one of the three actions, but the structure remains the same. As a result, the following sections will describe each action for every analysed VoQ. For the moment, this work covers only *growl* type voices, which are discussed below.

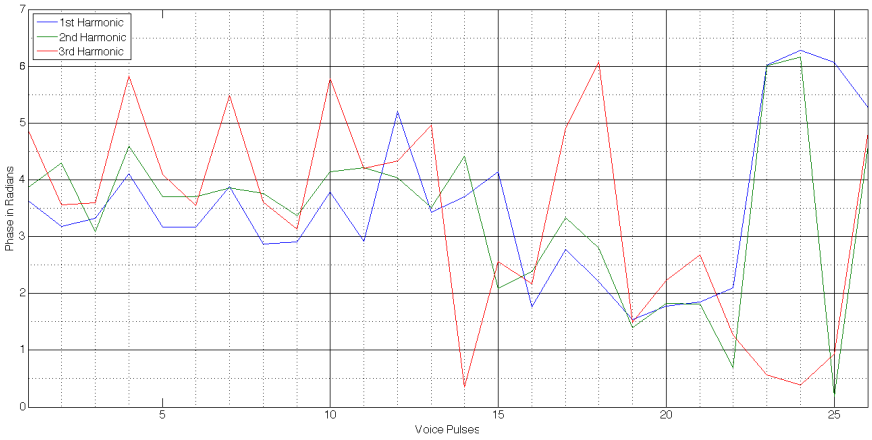
4.2 GROWL VOICES

4.2.1 *Definition*

In contrast to modal voices, which tend to exploit relatively stable harmonic time series and low energy level, *growl* voices can be described by rapid changes of timbre, timing and energy of source excitation events [7]. These changes result into appearance of sub-harmonics in the spectrum [20] and oscillations patterns in the time domain [7]. The following Figure 4, demonstrates these oscillations of the first three harmonics, for initial voice pulses after performing the WBHSM analysis to a representative *growl* utterance.



(a) Amplitude oscillations



(b) Phase oscillations

Figure 4: Oscillations derived from WBHSM

4.2.2 *Oscillation Modelling*

Observing the above figure, specific periodic patterns of oscillations can be easily derived. In addition to this, there are small voice pulse shifts (delay), regarding each harmonic. Thus, the main goal should be modelling these specific qualities that are describing the *growl* voice.

To do so, different approaches could be considered. The most straightforward ones, are voice conversion and spectral morphing based ones, described in [6, 7] respectively. While none of these methods exploit voice qualities characteristics, it is convenient to model the described oscillations with sinusoids, from extracting features of the cepstrum or modulation spectra.

To do so, let us first define $\overline{X}(m,k)$ as complex valued spectrum acquired from WBHSM, with magnitude $M(m,k) = |\overline{X}(m,k)|$ and phase $\Theta_(m,k) = \angle\overline{X}(m,k)$, where m, k denote harmonic bins and voice pulses respectively. Assuming that a modal voice can be characterised by stable harmonic series and low level noise energy [7], to obtain the residual part containing the oscillation patterns, we smooth $M(m,k)$ with an moving average filter yielding $M(m,k)_{harm}$. The residual part is recovered by :

$$M(m,k)_{res} = M(m,k) - M(m,k)_{harm} \tag{6}$$

Then for each one of the harmonic index in $M(m,k)_{res}$ and $\Theta_(m,k)$ we perform STFT analysis :

$$\overline{M}(k)_{res} = M(k)_{res} w(k) \tag{7}$$

$$\overline{M}(k)_{osc} = \sum_{k=0}^{N-1} \overline{M}(k) e^{-j2\pi \frac{kn}{N}} \tag{8}$$

where, $w(k)$ is the analysis window of length N derived by the desired number of voice pulses. Finally we define its magnitude and phase content as :

$$Osc(k)_{mag} = |\overline{M}(k)_{osc}| \tag{9}$$

$$Osc(k)_{phase} = \angle\overline{M}(k)_{osc} \tag{10}$$

As for the actual frequency values, they are computed by solving a 2nd order polynomial function, with its coefficients fitting $\max_{Osc(k)_{mag} \leqslant \frac{N}{2}}$. Moreover, it is vital to inherit the phase error of the above computation by:

$$D_{OSCphase}(k) = Osc(k)_{phase} - error \tag{11}$$

$$error = \mod (Osc(k)_{phase} - ph_{prp}(k) + \pi, 2\pi) - \pi \tag{12}$$

$$ph_{prp}(k) = \frac{Osc(k)_{phase} - 2\pi f}{N + \sum 2\pi \frac{f}{ws}}, \qquad \forall f \in \Re \tag{13}$$

where, f is a value derived from the 2nd order polynomial solution, contained in an arbitrary vector of frequencies $F(m, k)$.

And the oscillations for each harmonic m can be re-synthesised by the following formula :

$$Osc(m, k) = \cos(2\pi \sum_{k=1}^{N} Osc(m, k)_{mag} F(m, k) + D_{OSCphase}(m, k)) \tag{14}$$

which is valid for both magnitude and phase information. At this stage, it has to be stated that since Eq. 14 uses also the index m, that denotes the harmonic series, it is straightforward to iterate the above equations for each harmonic bin m, with respect to voice pulse k, as described.

Finally, a second observation from the above Figure 4, is a specific repetitive pattern of delay in voice pulses, regarding each harmonic. Thus, in addition to the oscillation pattern we also compute the delay $Del(k)$ between harmonic series by a simple autocorrelation function. By filtering the Eq. 14 with $Del(k)$, we achieve to the desired emulated oscillation.

4.2.3 *ANN Entrainment*

Apart from the methods incorporating direct transformations based on FD information, probabilistic approaches also do co-exist. In the latter case, an "acoustic domain" is being defined by features such as spectral envelope [1], which later are being fitted to a statistical model.

State of the art models for the described task are, Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). Unfortunately both of the described procedures are encountering problems in converging to a solution, because of restricted training data set [52]. On the other hand, GMM tend to over-fit, in the training procedure, or require over-smoothing of the imported data [53].

To this end, Artificial Neural Networks have been proved to perform as good as the above methods, overcoming the described problems [54]. Moreover, assuming that we are constrained with multidimensional data, regarding m harmonic series and k voice pulses for three parameters $D_{OSCphase}(m,k)$, $F(m,k)$ and $Osc(m,k)_{mag}$ describing magnitude $M(m,k)_{res}$ and phase $\Theta_(m,k)$ oscillations, two approaches have been proposed.

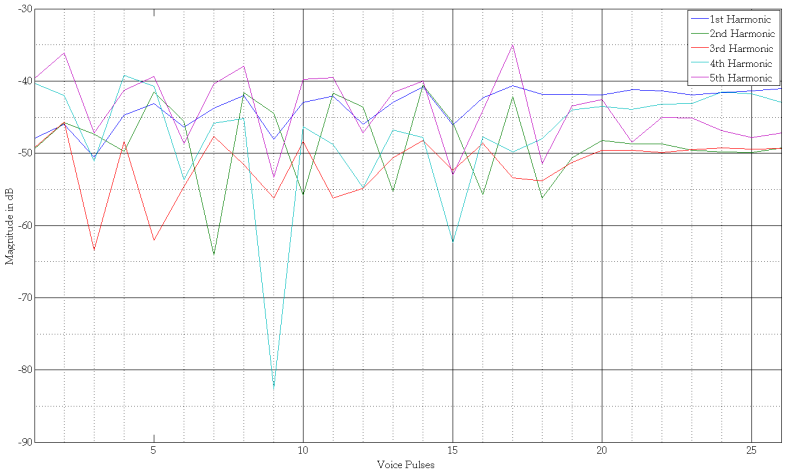
The first one, encapsulates radial basis functions, in order to converge to an optimal solution [55]. A reported issue of these functions is the ill-conditioned or dense output matrix, which collocates the output variables, meaning that in our intention to predict the time series evolution of each parameter controlling the oscillation synthesis formula, the time evolution of the *growl* will be smeared.

Aiming for the most well-fitted scenario of VoQ modelling, where the time evolution can enrich the perceived naturalness and describe best the actual phenomena, we introduce the a set of ANNs, denoted as ANN ensemble, which is based on the approach described in [56]. The general idea is to acquire a generalised model that can actually predict the time evolution of the parameters depending on a minimum amount of input values, without smearing the initial information. As a result, 8 ANNs were trained for predicting the aforementioned parameters (*magnitude and phase oscillation frequency(rate), amplitude(width), phase and harmonic group delay*).

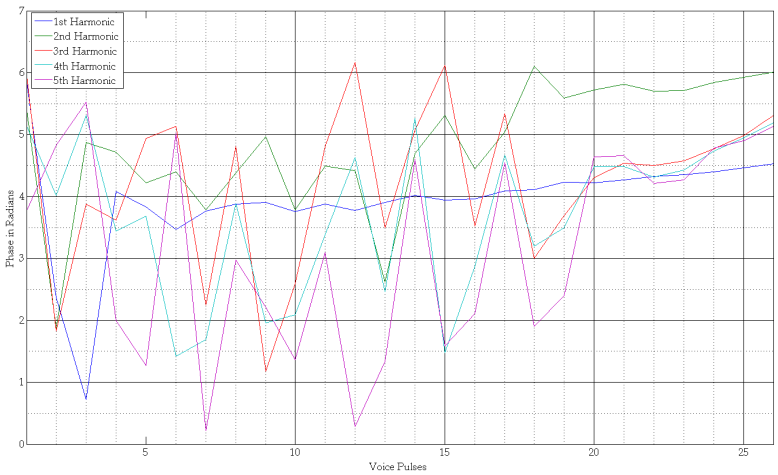
4.3 CREAKY VOICES

4.3.1 *Definition & Methodology*

Creaky voice is a another type of rough voices, which normally can be manifested among with local and less occurrent irregularities [11]. Perceptually from a transmission, the listener can actually perceive the separate vocal fold vibrations, with main characteristics such as short boosts and abrupt changes in amplitude and frequency [10].



(a) Amplitude oscillations



(b) Phase oscillations

Figure 5: Oscillations derived from WBHSM

These type of voices are produced by thick, less distance separated vocal cords, or when the folds above the vocal cords grasp the activated vocal cords and dampening a portion of the vocal cord vibration [10, 11]. Moreover, in [10] short glottal pulses occurring at low fundamental frequencies and combined with a “double-periodicity” phenomena, caused by diplophonic irregularities in the fundamental period, were reported.

Analysing such a voice using the described methodologies of WBHSM, we can define an acoustic feature space containing magnitude and phase information, throughout useful frames that merge to voice pulses. An illustration can be previewed in Figure 5.

The illustration above, demonstrates the aforementioned phenomena derived of such an analysis upon a recorded voice utterance. Again, the most representative voice pulses and harmonics are being presented.

As it can been seen, the most prominent effects are the magnitude and phase oscillations, which this time have a ceased effect across the voice pulses and most of the peaks are synchronised. Meaning that, only at the beginning of the utterance a critical pattern occurs, which is being "smoothed" along time and delay pattern computations will be insignificant at the next stage of transformation.

The last statement, can be exploited by two methodologies. The first one is to manipulate the oscillation width, rate and phase parameters of the entrained *growl* ANN model, in order to achieve the desired output oscillation pattern. Simply, by linearly decreasing the oscillation width and interpolating the adequate rate, while neglecting the group delay information, almost perceptually similar effects can be achieved.

Aiming for natural transformations and phenomenal replication and since the analysis part has to be done once and ANN’s are covering the part of prediction, a second strategy was followed. which is the same of analysis, parameterisation and learning of *growl* voices. This strategy includes the following routine :

1. Define an audio corpus including creaky voices’ utterances.
2. For each recorded file, perform the WBHSM.
3. Define the desired number of harmonics for analyses for magnitude and phase information.
4. For each harmonic follow Equations [6 — 13].

5. Re-synthesise the oscillations using Equation [14].

Moreover in [10], it is claimed that there is a probability of higher harmonic energy increase, due to narrow glottal pulses at the specific period, when the upper folds are dampening active vocal cords, resulting into a relatively flat spectrum with specific steepness in terms of spectral tilt [10]. By analysing the "in-house" sub-data set containing *creaky* voices, such phenomena were not observed. Assuming the above probability of existence, we also have considered adding such a feature that will allow the boosting of higher harmonics, which will be applied after the synthesis and application of oscillation parameters. The next subsection, is concerned with the overview of the method followed for boosting specific spectral regions.

4.3.2 *Spectral Harmonic Enhancement*

The following operation can be described as a high shelving equaliser, that can be found in typical spectral processors, with variable frequency. The reason for embedding variable frequency inside the procedure is that all human voice utterances are characterised by specific regions in the magnitude spectrum containing useful information [58]. This information is very important and sensible in human perception and is consisted by speaker dependent spectral slopes and shapes [58].

For describing such spectral shapes different approaches exist. The most convenient with extra care given to emotional features of humans voice is being presented in [58]. The selection of this approach, was based on the correlation and overlapping spectral information handling between emotional speech and VoQs. Moreover, speaker dependent analysis based on fundamental frequency has been proven to perform better [58].

It has to be stated, that for this task we are focusing on the results of the aforementioned research, regarding the selection of the "convenient" position in spectrum, henceforth called "pivot", in order to perform our equalisation instead of estimating or extracting features proposed by [58]. Where in wide-band conditions, is even more simple, assuming that each harmonic is allocated one spectral bin.

Towards this notion, the aim is to localise a specific pivot which splits the spectrum into two parts that contain major information. More specifically, the first and lower part of the spectrum information related to phonetics enabling

the characterisation of different vowels, whereas the second part conveys information about VoQs [58]. As far it concerns the location, it is based on the selection of a multiple of fundamental frequency. Results in [58], showed that the usage of 10th harmonic will practically enforce this separation.

After partitioning these two spectrum components, the user can boost the second part of the spectrum, which contains the higher harmonics, up to a desired value in dB scale, which will not exceed the lower behalf. Finally, it should be mentioned that this operation is being implemented after the oscillation resynthesis, derived from Equation [14].

5

EXPERIMENTAL PROCEDURE

5.1 DESCRIPTION

In this Chapter 5, two main objects are discussed. The first one is related to the validation of the above, proposed methodology described in Chapter 4, providing information for the utilised audio corpus, analyses and re-synthesis parameters and of course the structure of the ANN ensemble, for each VoQ. The last one, is concerned with a user-based evaluation by conducting listening tests where, subjects have to rate a set of processed, by the proposed method, singing voices.

5.2 PERFORMANCE ASSESSMENT

5.2.1 *The Growl Audio Corpus*

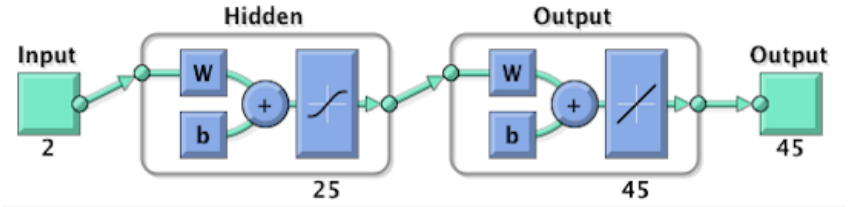
In order to assess the performance of our proposed methodology, a sub-data set of 14 "in-house" *growl* recordings was defined. This set, contained a variety of *growl* utterances, by a specific male singer, dissimilar in time length and pitch.

Then, each one recording was analysed with the WBHSM algorithm and oscillation features were extracted. Finally, these features were fed into target observations of 8 ANNs. The number of harmonic series m was set to 45, according to the practical observations of the significant energy fluctuation throughout the audio corpus.

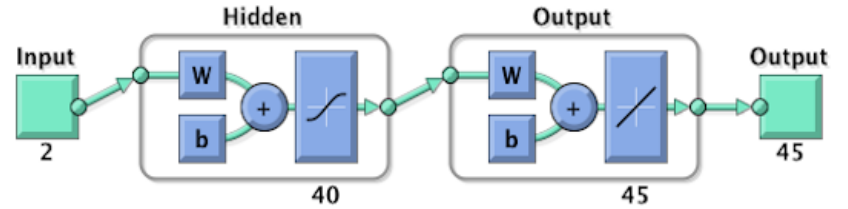
As far it concerns the input values, the fundamental frequency of the voice utterances and the manually annotated time stamps, describing the length and the start/ending points of the utterance, were used.

Focusing on the ANN structure, three main categories of Feed-Forward Neural Networks were used among the 8 individual ones. More specifically, for the oscillation width and rate "Category A"' was used, while phase incorporates

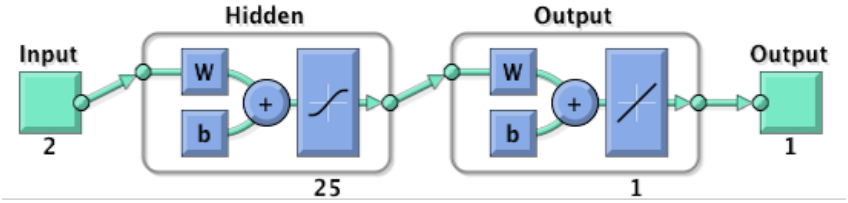
"Category B" and harmonic delay patterns employ "Category C", as illustrated in the following Figure 6 with the corresponding parameters in Table 2.



(a) Category A



(b) Category B



(c) Category C

Figure 6: ANN Categories Used

Table 2: ANN Parameters

Derivative Function	Static Derivative
Data Division	Random
Learning Function	Levenberg-Marquardt back-propagation
Performance Function	MSE
Error Normalisation	Active
Number of Layers	2
Number of Layers	2
Number of neurons	Variable (see Figure 6)
Initialisation Function	Layer-by-layer network initialisation

After the successful entrainment, each imported voice signal is analysed with WBHSM algorithm, the user selects time segments for the voice transformation and fundamental frequency values alongside with an arbitrary time stamp is activating each one of the 8 artificial networks, that are predicting the time evolution of the modelled parameters. Then, the predicted values are being given to the synthesis formula Eq. 14, filtered by the predicted time delay and simply added to the harmonic amplitude and phase information. Finally, the transformed voice is being recovered into TD, using the re-synthesis part of the WBHSM algorithm. An illustration of the transformation procedure is being given in Figure 7.

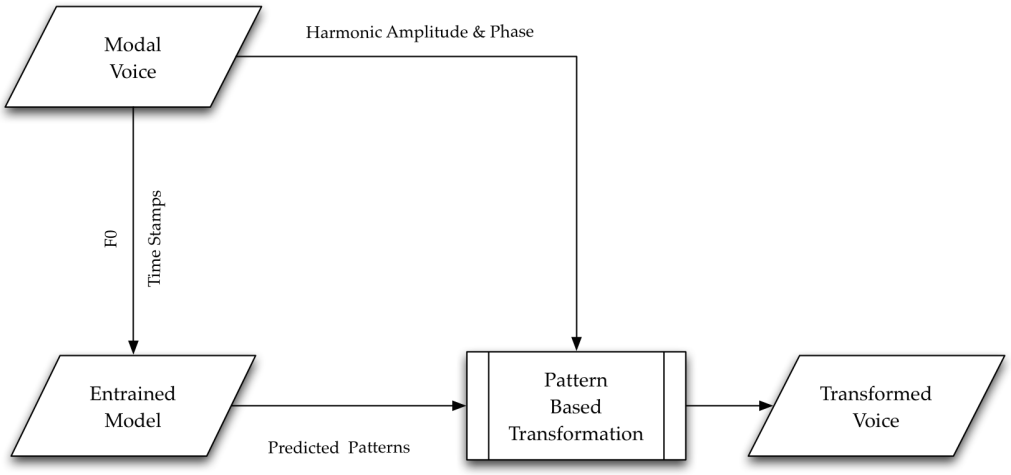
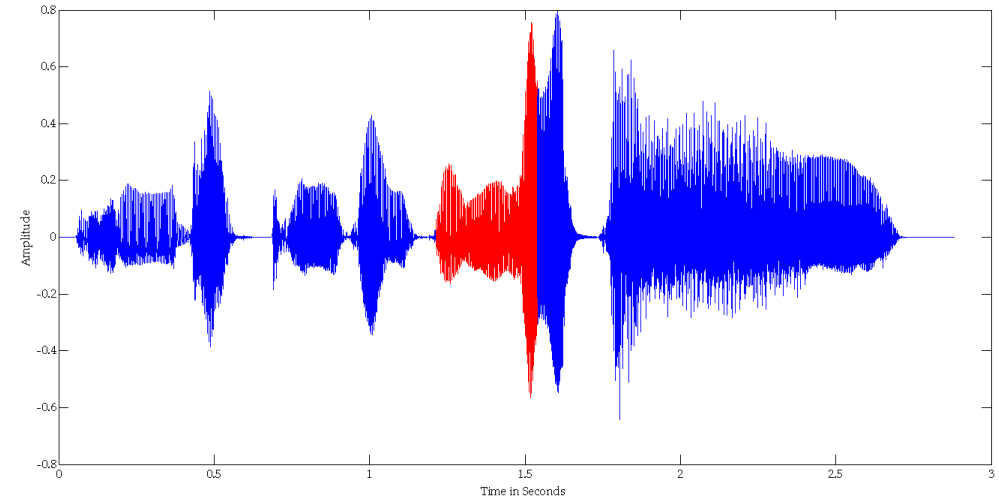


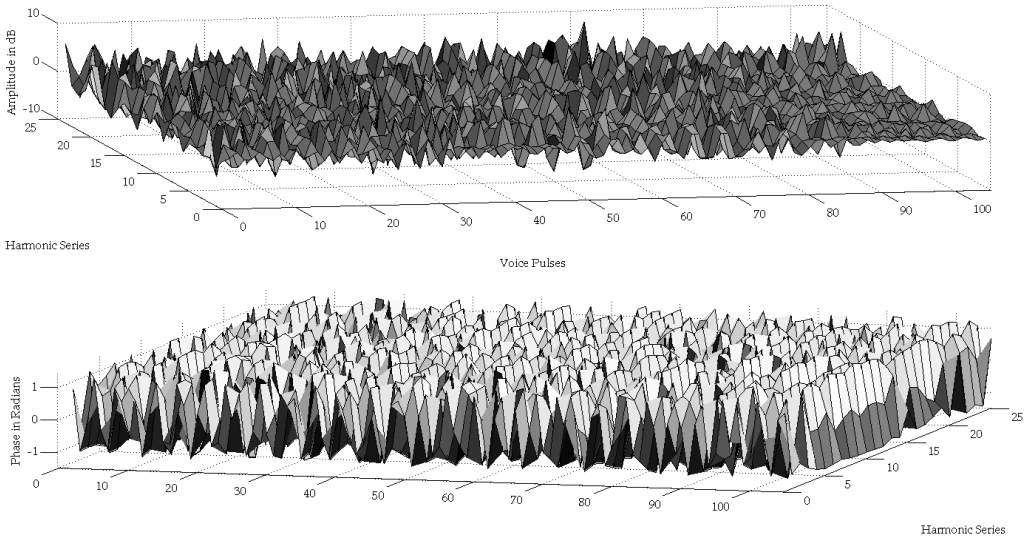
Figure 7: Transformation Stage

For validating the above procedure two female and two male singing voice excerpts, were processed by the described system [1]. The voice transformation was applied to selected voiced regions, marked with red colour in the following Figures, where their corresponding time domain signal representations among the predicted oscillations are being demonstrated below.

1 An audio demonstration, using these excerpts, will be given in the presentation

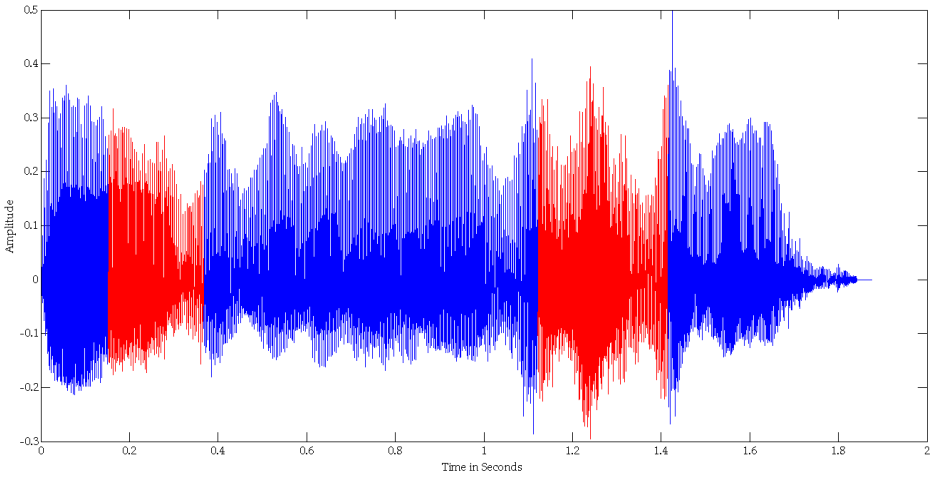


(a) Time Domain Signal

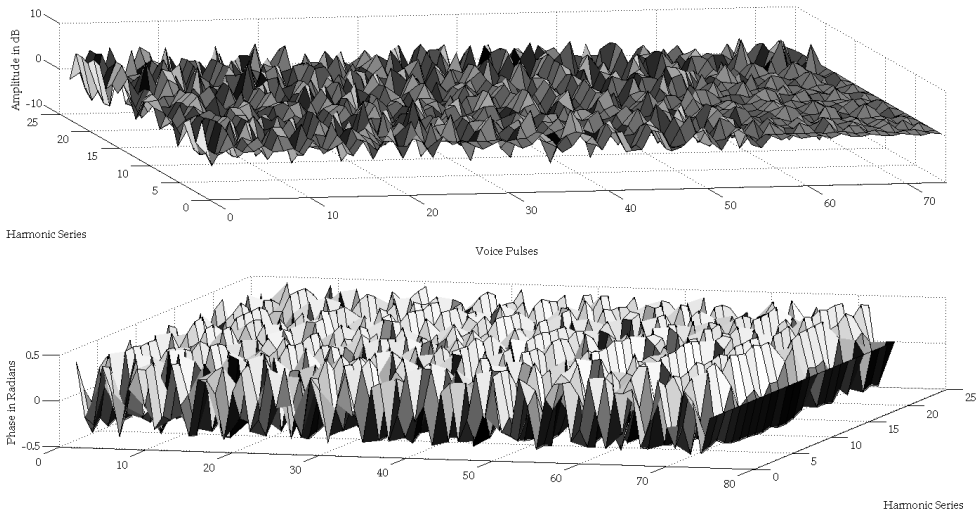


(b) Predicted Amplitude & Phase Oscillations of 25 out of 45 Affected Harmonics

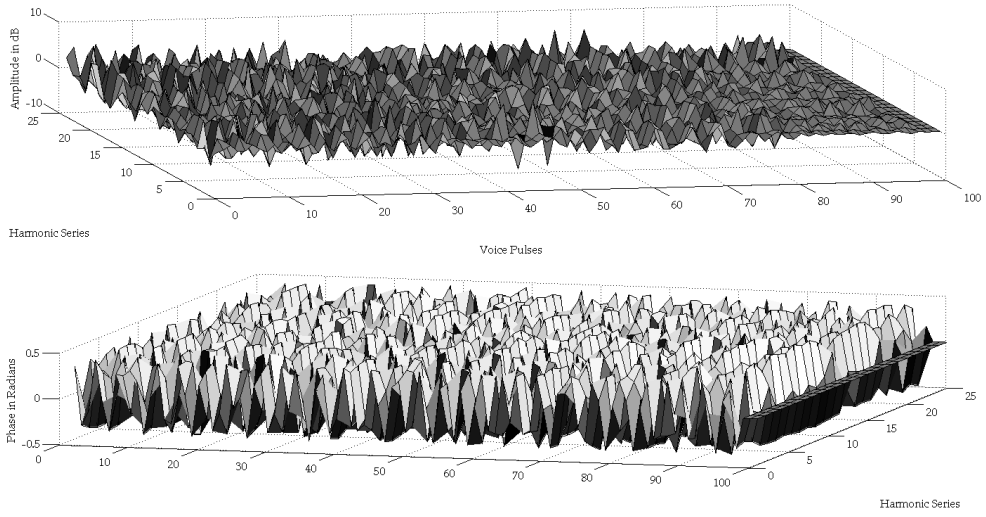
Figure 8: Female A Excerpt



(a) Time Domain Signal

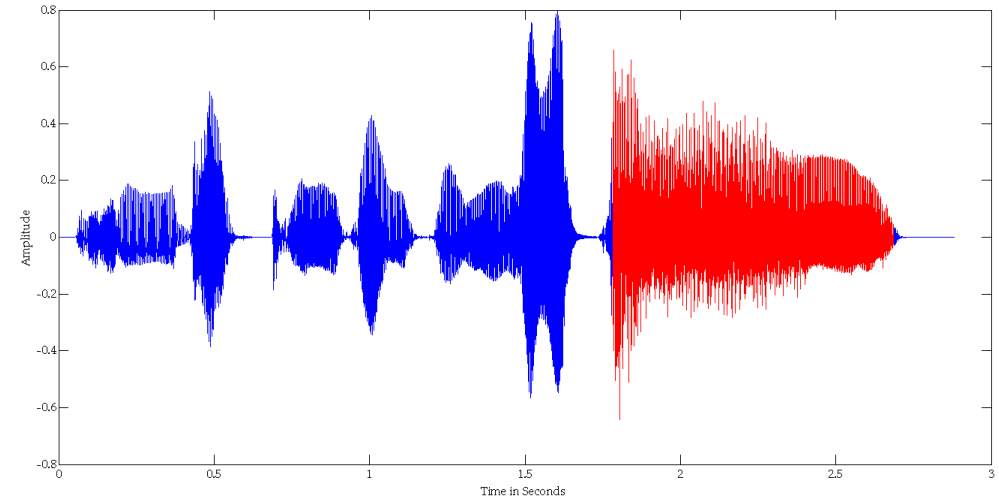


(b) Predicted Amplitude & Phase Oscillations of 25 out of 45 Affected Harmonics

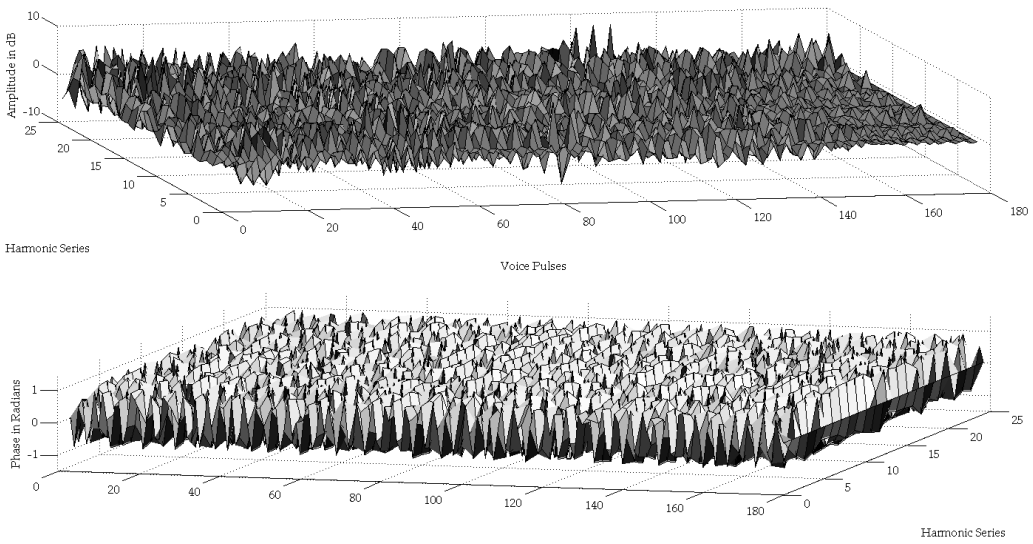


(c) Predicted Amplitude & Phase Oscillations of 25 out of 45 Affected Harmonics

Figure 9: Female B Excerpt

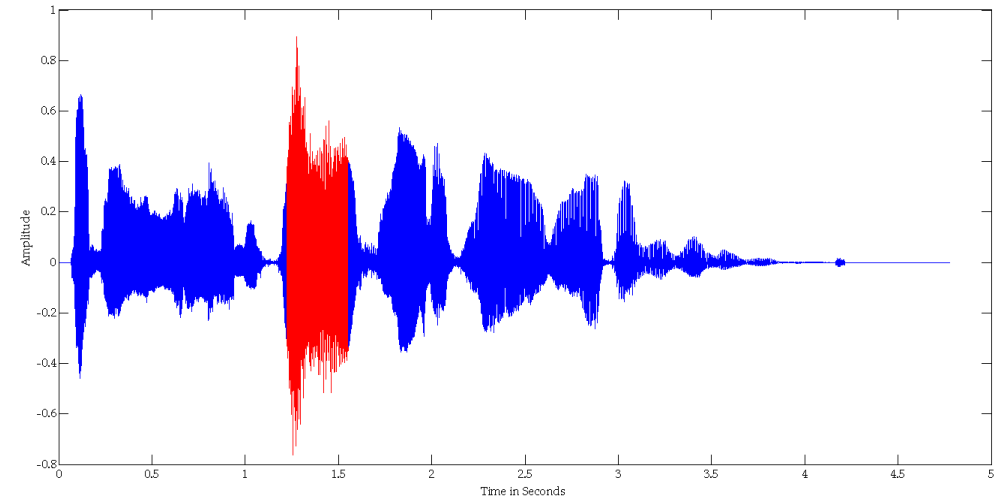


(a) Time Domain Signal

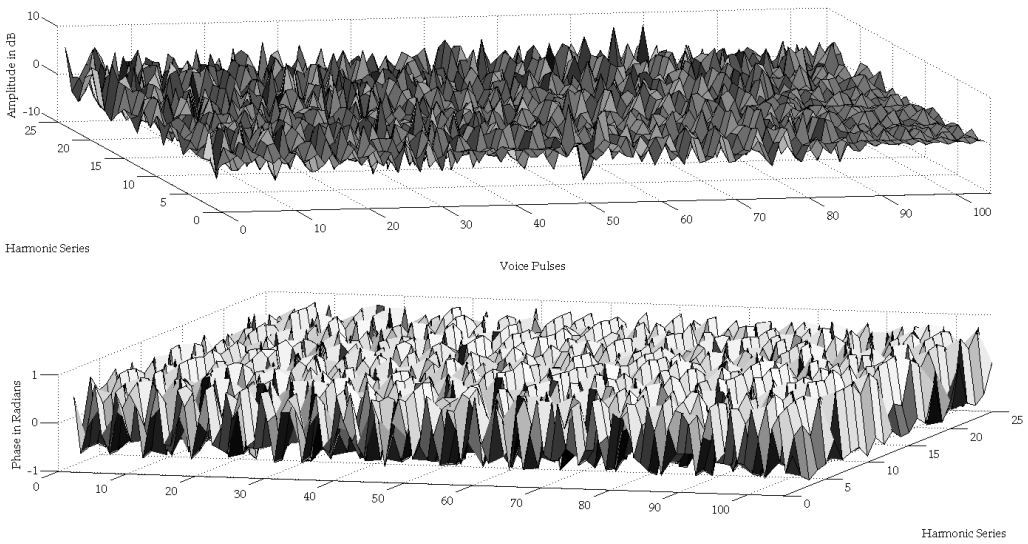


(b) Predicted Amplitude & Phase Oscillations of 25 out of 45 Affected Harmonics

Figure 10: Male A Excerpt



(a) Time Domain Signal



(b) Amplitude & Phase Oscillations of 25 out of 45 Affected Harmonics

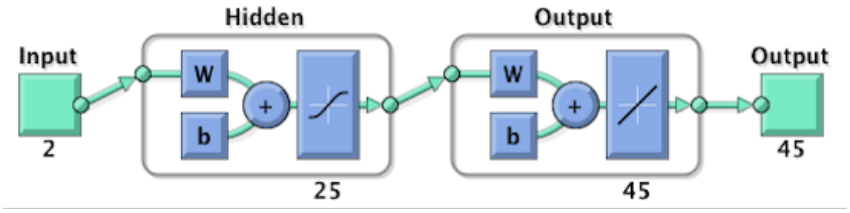
Figure 11: Male B Excerpt

5.2.2 *The Creaky Audio Corpus*

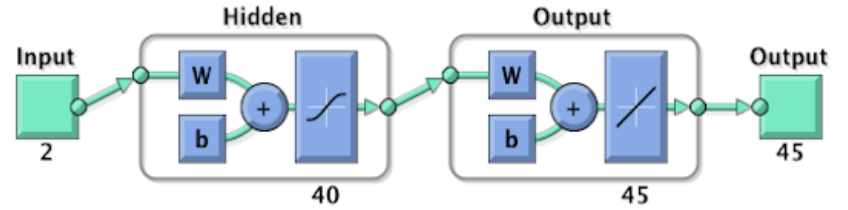
Another sub-data set of 10 "in-house" recordings containing *creaky* voices was defined. Similarly to the *growl* one, it contained a variety of *creaky* voice utterances, by a specific male singer, dissimilar in time length and pitch.

Each one recording was analysed with the WBHSM algorithm and the aforementioned oscillation features were extracted. Finally, these features were fed into target observations of 6 ANNs. The number of harmonic series m was set to 45, for preserving a homogenous module for the whole system.

Once again the input values, were the fundamental frequency of the voice utterances and the manually annotated time stamps, describing the length and the start/ending points of the utterance. Finally, the same structure of ANN's was incorporated, with the only difference that 'Category C" was not used, for this specific routine, as illustrated in the following Figure 12 with the corresponding parameters in Table 3.



(a) Category A



(b) Category B

Figure 12: ANN Categories ussed for *creaky* voices

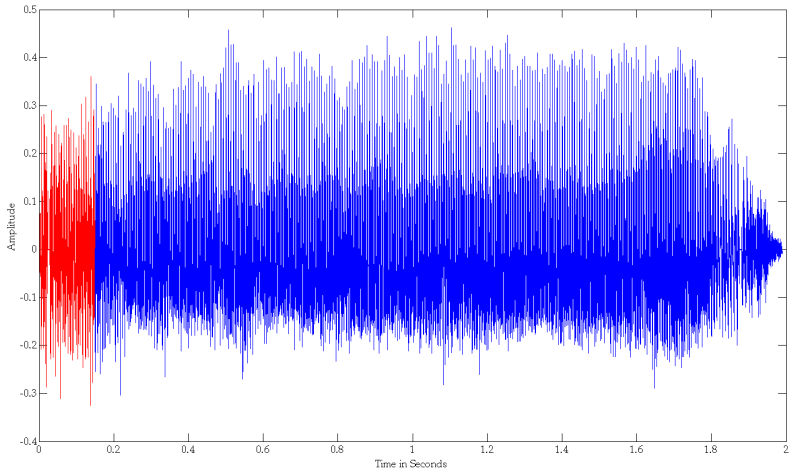
Table 3: ANN Parameters ussed for *creaky* voices

Derivative Function	Static Derivative
Data Division	Random
Learning Function	Levenberg-Marquardt back-propagation
Performance Function	MSE
Error Normalisation	Active
Number of Layers	2
Number of Layers	2
Number of neurons	Variable (see Figure 12)
Initialisation Function	Layer-by-layer network initialisation

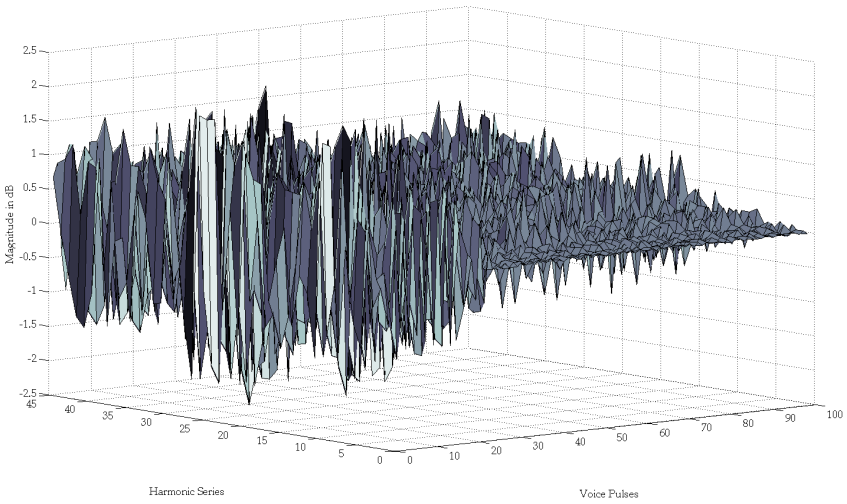
After the successful entrainment, the same approach as *growl* voices is followed, demonstrated in Figure 7. At this stage, it has to be stated that the Group delay procedure is being neglected from this strategy.

For validating the followed strategy two male modal voice excerpts, were processed by the described system. The main reason behind the usage of modal voices, is that *creaky* voices represent mainly voice disorders that are not frequently observed into singing voice styles or in the case that they are existent, no modal excerpt is available for performing the comparison.

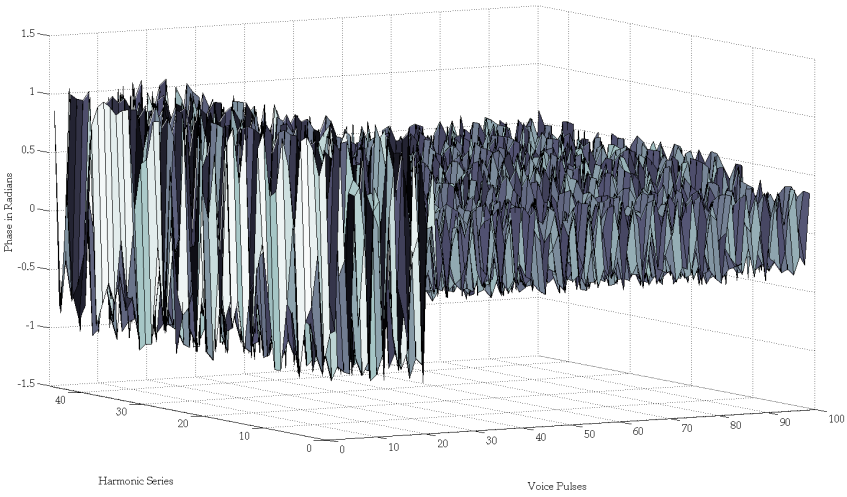
The following figures demonstrate the predicted oscillations, as an outcome of the ANN ensemble entrained with *creaky* audio corpus. As an input, two male voice excerpts were used and the same routine as *growl* voices was performed. It has to be stated, that only 25 out of 45 affected harmonics are being displayed for providing a more coherent overview.



(a) Time Domain Signal

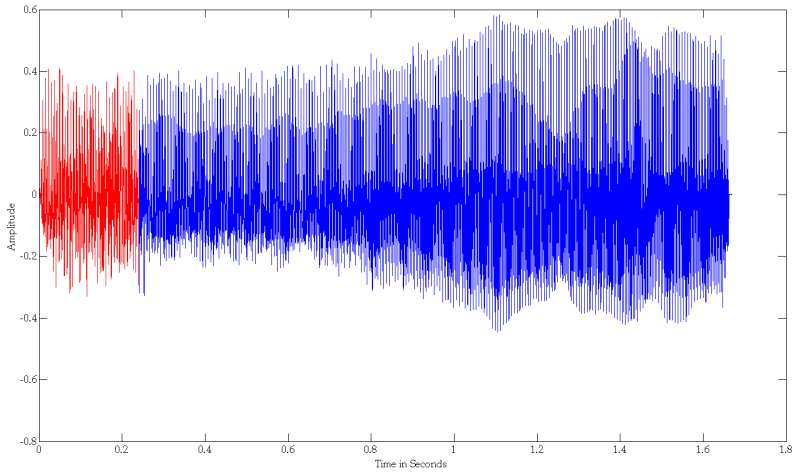


(b) Predicted Magnitude Oscillations

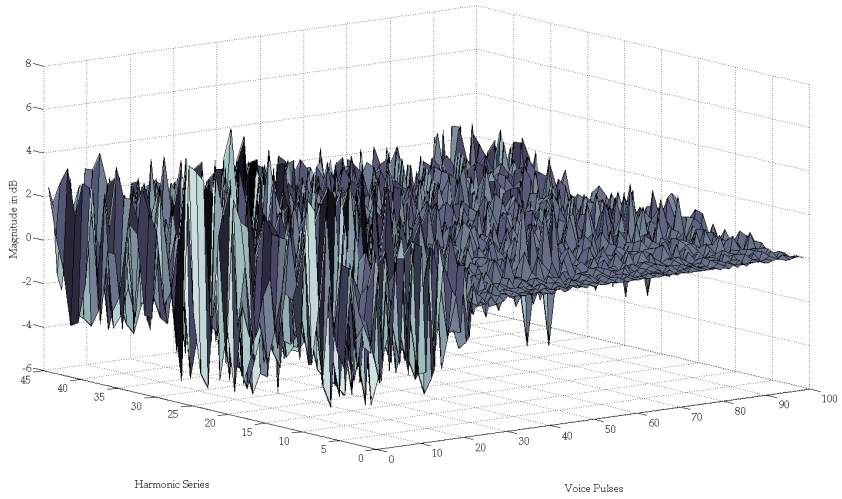


(c) Predicted Phase Oscillations

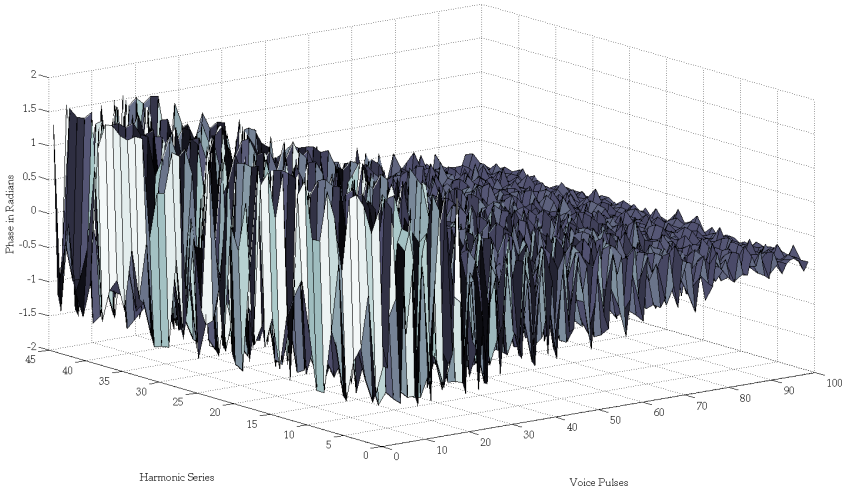
Figure 13: Male A Excerpt



(a) Time Domain Signal



(b) Predicted Magnitude Oscillations



(c) Predicted Phase Oscillations

Figure 14: Male B Excerpt

5.3 SUBJECTIVE EVALUATION

As a final stage of evaluation, a series of listening tests is programmed to be performed. For this task singers and musicians that are involved professionally into the recording, mixing and producing stages will be considered.

In addition to this, a Graphical User Interface (GUI) in Matlab [57] has been developed to carry out the experimental procedure. Each user shall listen to a "pseudo-random" sequence, including the original and the transformed voice excerpt, while the users shall rate the perceived naturalness and expressivity of each excerpt. The overview of the developed GUI is being given in Figure 15.

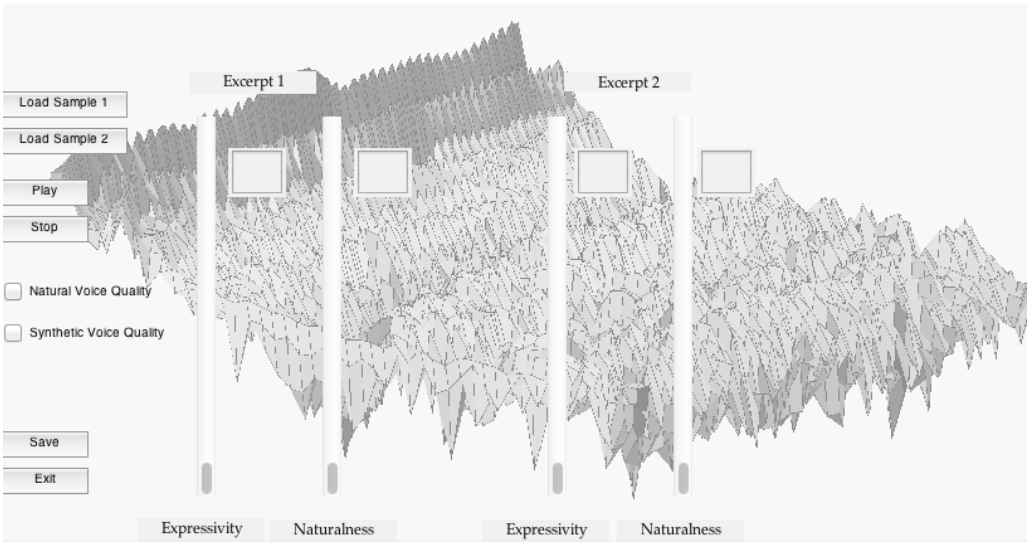


Figure 15: Subjective Assessment GUI

As for the actual listening tests, 8 audio professionals, with background in audio mixing, production and of course familiar with singing voice styles incorporating *growl, etc.*, were selected with main task to rate each excerpt. The rating was based on perceived "Expressivity", in case of multitrack processed corpus, and perceived "Naturalness" plus denoting if they believed that the excerpt contained a synthesised or natural VoQ, in case of singing voice excerpts.

Thus, the listening tests can be divided into two categories. In the first one, where the *growl* transformations are included, the approach of evaluation is based on the perceived "expressivity", evaluated through processed multitrack audio sessions, and on perceived "naturalness", evaluated through small singing voice excerpts. The latter category, contains the *creaky* voice transformations

and the selected approach of evaluation is the comparison of natural and synthesised, by the system, voices, where the user has to denote the natural and the synthetic one.

The main reason behind such an approach, is the initial motivation to evaluate the system in such cases where not only quality matters, but also as satisfactory performance in an production chain, since *growl* voices are being denoted in popular music genres. On the other hand, *creaky* voices have a more contained impact, regarding their occurrence and availability in music scene or recordings.

The apparatus used in this experiment can be previewed in Table 4. As far it concerns the stimuli, 4 singing voice excerpts and 2 well-known songs' multi tracks were processed by the proposed system, using the *growl* trained module and 4 other modal voices processed using the *creaky* trained module.

Table 4: Experimental Audio Apparatus

Electroacoustic Device	AudioTechnica ATH-M40FS Studio Headphones
ADC Converter	Native Instruments Komplete Audio 6
Playback Software	MATLAB [57]
Relevant Hardware	Apple MacBook Pro i7 2.66 GHz 15"
Relevant Software	OSX 10.9.4, CoreAudio

The algorithm followed, for the subjective assessment, is described in the following steps :

1. Each subject was exposed to the aforementioned GUI after a brief explanation of the procedure and the goals of this operation.
2. A desired sound pressure level was set, by modifying level parameters of the apparatus in Table 4.
3. The short-length singing voice excerpts were loaded and the subject pressed "Play" at the desired time instance.
4. By pressing "Play" automatically, the audio samples were normalised and pseudo-randomly placed in an order.
5. After the stimuli, the user can listen that sequence, as many times as desired.

6. Upon arbitrary decision, the subject evaluates the perceived “Naturalness” of each segment.
7. Then, it has to be denoted, if the utterance, containing each VoQ was synthetic or natural.
8. Up until all short-length excepts are over repeat the steps above.
9. Else, load the samples containing the processed multitrack mixes.
10. This time, the subject rates only the perceived “Expressivity”.
11. In case *creaky* voices are evaluated, the subject rates if he perceived a “Natural” or “Synthetic” VoQ.

6

EXPERIMENTAL RESULTS & DISCUSSION

As far it concerns the results of the subjective assessment ,that encapsulated a series of listening tests with audio professionals, the following Figures 16 - 18, demonstrate the performance of the proposed system. More specifically, the three-way assessment (perceived naturalness, expressivity and distinction/-transparency) aims to explore the equilibrium between "benefits" and "pitfalls". In other words, it denotes the trade-off of what perceptual characteristic you have to sacrifice in order to gain another one. The third dimension of assessment, evaluates the system in terms of transparency, embedding a VoQ that is being widely used in emotion speech recognition or TTS tasks.

At this point it has to be stated that a grade of 10 means that perceptually an excerpt sounds natural and 1 purely "synthetic" or un-natural. While the same grading system was followed for expressivity purposes.

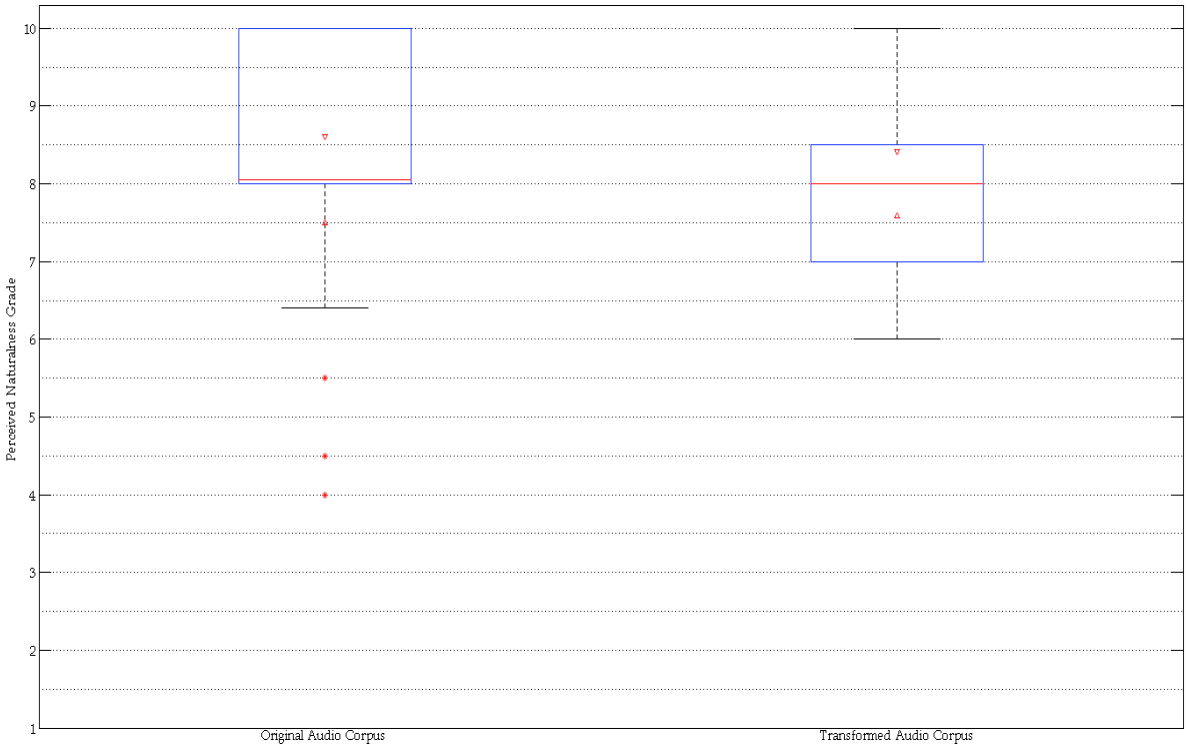


Figure 16: Perceptual Naturalness Grade

Focusing on the actual results and the constrain of perceived naturalness, in terms of how "artificial" or "synthetic" this operation may sound, the transformed audio corpus achieved almost adequate median grade (8) to the original / non-transformed audio corpus (8.1). In addition to this, "Tukey box-whisker" diagrams also demonstrate the outlier grades that the original audio received, which are denoted with red asterisks. This can be interpreted as a disagreement between the subjects, where not always the original audio can provide an arbitrary perceptual naturalness. On the other hand, the "compactness" of the transformed audio corpus box-plot, shows that most of the subjects agree on a less diverse span of the aforementioned perceptual grade, but still high enough to compete the former one.

By sacrificing a small proportion of perceived naturalness, it was observed that the perceptual expressivity, in terms of singing voice performance, can significantly be increased, using the aforementioned procedure. The following Figure 17 demonstrates the aforementioned increase, where even in the outliers an increase of 0.3 was achieved. The interesting fact is that the increment grows for the median values, achieving a divergence of 1.4 between original and transformed audio excerpts.

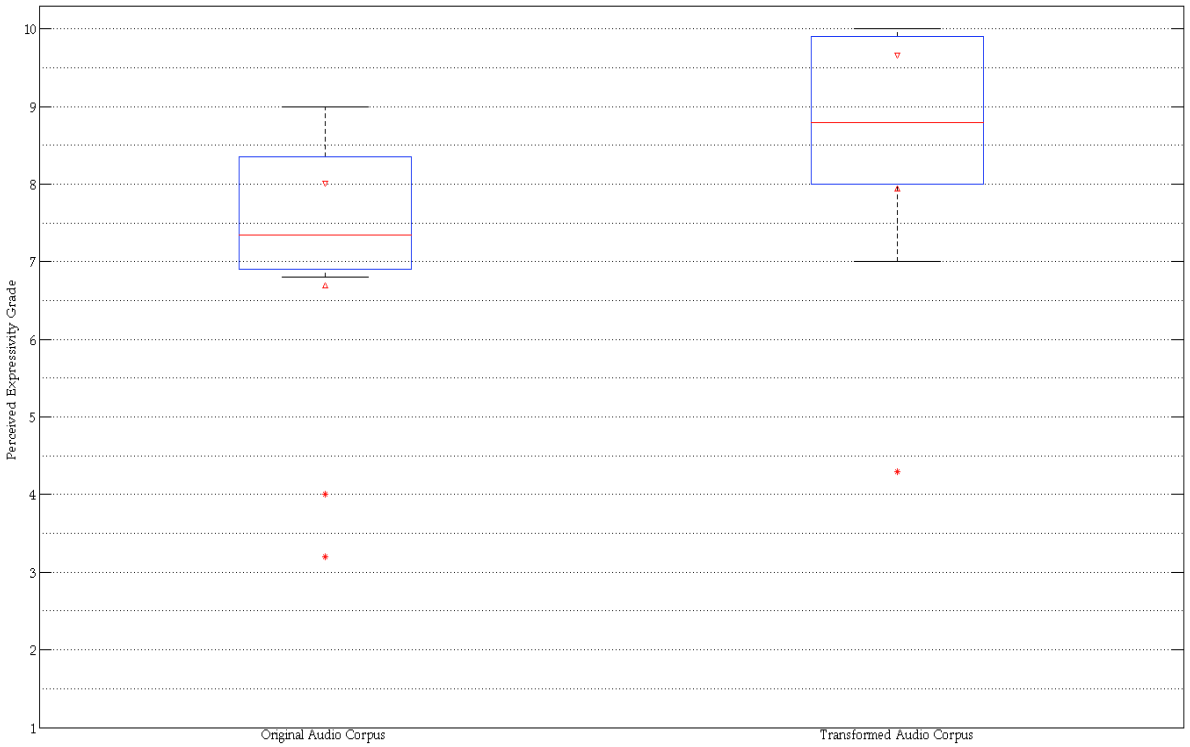


Figure 17: Perceptual Expressivity Grade

When it came for the subjects to distinct which could be the transformed or "synthetic" audio excerpt, using the *creaky* based transformed examples, the success rate of denoting the original recordings is up to 50%. On the other hand, $\approx$ 55.5 %, of the transformed audio corpus was classified as natural or non-synthetic audio excerpt.

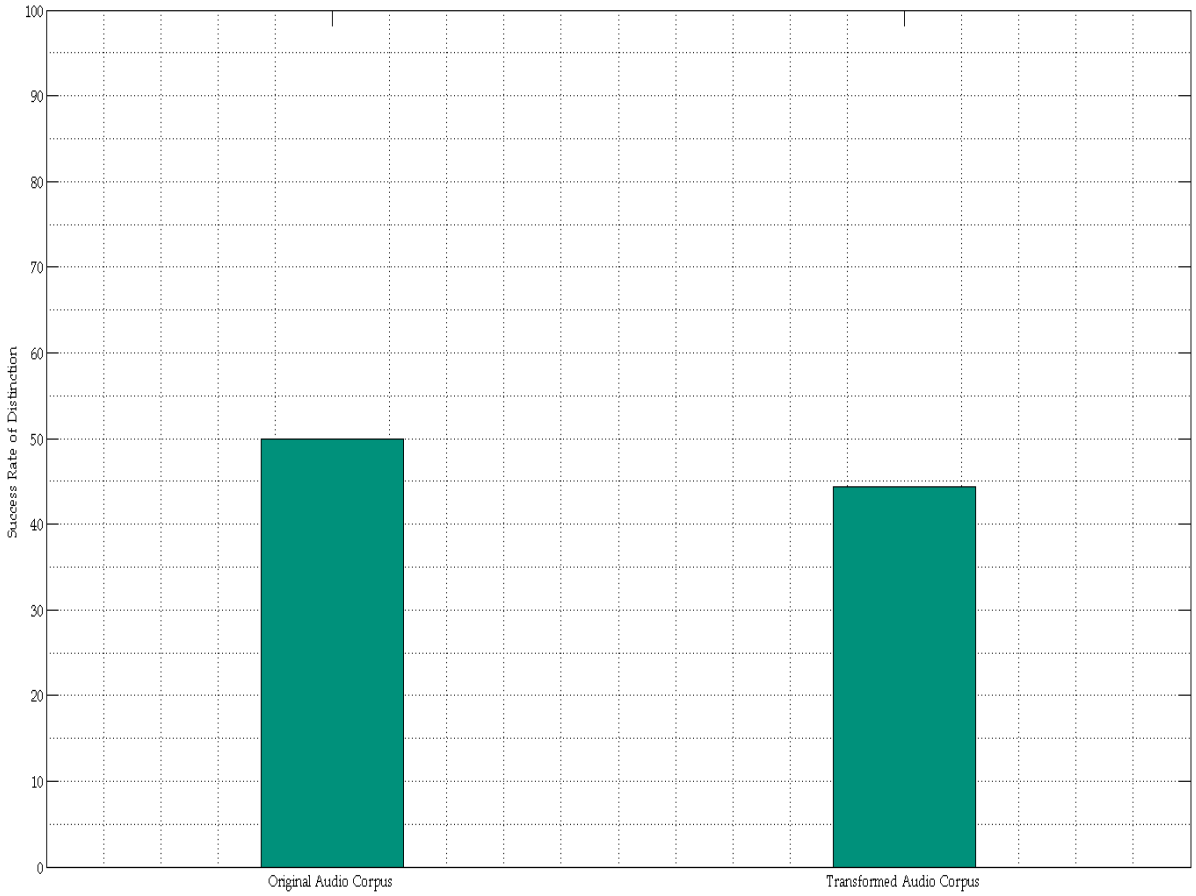


Figure 18: Perceptual Expressivity Grade

Finally, it has to be stated that the above results were achieved using the default values of re-synthesis and transformation, which were the same for all audio excerpts. Thus, it is highly prominent that specific parameterisation can yield even better results.

7

CONCLUSIONS & FUTURE WORK

In this work, a system capable of generating and adapting transformation parameters used in voice processing was presented. Briefly, it combines different modules that are responsible for specific tasks such as, robust voice analysis and re-synthesis (namely, Wide-Band Harmonic Sinusoidal Modelling), voice pattern/feature analysis and extraction, "learning" incorporating Artificial Neural Networks ensembles and last but not least predictive voice transformations based on the activation of Neural Networks. The whole procedure stands for modelling and synthesising different phenomena, occurring to different voice qualities, that allow the transformation of input voices.

In addition to this, the proposed method overcomes drawbacks and limitations of previous approaches and also performs with the least amount of prior knowledge of an input signal (fundamental frequency and arbitrary desired time length of transformation) or audio-corpus restrictions. Moreover, we exploit a deeper analysis of these phenomena which enables intelligent and adaptive processing stages, meaning that can be incorporated in any adaptive/intelligent digital audio processing module [59, 60].

Listening tests evaluating the performance of the proposed system, showed that the transformations yield perceptually enhanced expressivity, in terms of a singing voice performance, without risking any critical loss of the perceived naturalness, as a whole operation. In addition to this, when it comes to distinction tasks, regarding transformation of *modal* voice excepts, the perceptual audible transparency is high enough for "synthetic" audio examples to smoothly blend with natural ones.

Finally, future work could be focused towards these extensions :

- Automated selection of the time instance, where a VoQ should transform the input voice.
- Assuming the above, it can also incorporate MIR tasks not only the time instance selection, but also for sophisticated parameters that control the oscillation re-syntheses.

- Examining spectral deviations from voice pulse to voice pulse (combining harmonic and time series prediction).
- Extending the analysis to adapt different VoQs that are not "time-dependent", using different representations.
- Integrating the entrainment part into pattern recognition or classification tasks.
- Embedding it into various voice effect processors accompanying harmonisation ones, as it was lately observed into new pop/rock productions.

REFERENCES

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion", in IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, March, 1998.

[2] C. Gobl, "A preliminary study of acoustic voice quality correlates", in Journal of STL-QPSR, vol. 30, no. 4., pp. 009-022, 1989.

[3] K.I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama, "Growl voice in ethnic and pop styles", in Proceedings of International Symposium on Musical Acoustics (ISMA 2004), Nara, Japan, April, 2004.

[4] M. Vasilakis, and Y. Stylianou, "New Trends in Voice Pathology Detection and Classification M & A of Vocal Emissions : Spectral jitter modelling and estimation", in Journal of Biomedical Signal Processing and Control, vol.4, no.3, pp. 183-193, July, 2009.

[5] J. Bonada, "Wide-band harmonic sinusoidal modelling", in Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September, 2008.

[6] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng and H. Li, "Exemplar-based Voice Conversion using Non-negative Spectrogram Deconvolution", in Proceedings of the 8th ISCA Speech Synthesis Workshop, 2013.

[7] J. Bonada, and M. Blaauw, "Generation of growl-type voice qualities by spectral morphing", in Proceedings of 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), Vancouver, Canada, May, 2013.

[8] D. Abercrombie, "Elements of General Phonetics", Edinburgh University Press, 1967.

[9] P. Ladefoged, "The Features of the Larynx", in Journal of Phonetics, vol. 1, pp. 73-83, 1973.

[10] D. H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", in Journal Acoustical Society of America, vol. 87(2), pp. 820-857, 1990.

[11] J. Laver, "The Phonetic Description of Voice Quality", Cambridge University Press, 1980.

[12] J. H. Esling, "Pharyngeal consonants and the aryepiglottic sphincter", in Journal International Phonetics Association, vol. 26(2), pp. 65-88, 1996.

[13] L. Fuks, B. Hammarberg, and J. Sundberg, "A self-sustained vocal-ventricular phonation mode: Acoustical, aerodynamic and glottographic evidences", KTH TMH-QPSR,3/1998, pp. 49-59, 1998.

[14] M. Lugger, and B. Yang, "Extracting voice quality contours using discrete hidden Markov models", in Proceedings of the Speech Prosody, Campinas, Brazil, May, 2008.

[15] R. Shrivastav, A. Camacho, S. Patel, and D. Eddins, "A model for the prediction of breathiness in vowels", in Journal Acoustical Society of America, vol. 129(3), pp, 1605-1615, 2011.

[16] C. Gobl, and A. Ni Chassaide, "The role of voice quality in communicating emotion, mood and attitude", in Journal Speech Communication, vol. 40, pp.189-212, 2003.

[17] C. Monzo, A. Calzada, I. Iriondo, and J.C. Socoro, "Expressive speech style transformation: Voice quality and prosody modification using a harmonic plus noise model", in Proceedings of Speech Prosody, no. 100985, Chicago, USA, 2010.

[18] R. Marxer, and J. Janer, "Modelling and separation of singing voice breathiness in polyphonic mixtures", in Proceedings of 16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland, September, 2013.

[19] B. H. Story, "Physical Modelling of Voice and Voice Quality", in Proceedings of VOQUAL'03, Geneva, Switzerland, August, 2003

[20] A. Loscos, and J. Bonada, "Emulating rough and growl voice in spectral domain", in Proceedings of 7th International Conference on Digital Audio Effects (DAFx'04), Naples, Italy, October, 2004.

[21] D. Mehta, and T. F. Quatieri, "Synthesis, analysis and pitch modification of the breathy vowel", in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 199-202, New Palz, New York, October, 2005.

[22] P. J. B. Jackson, and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech", in IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 713-726, 2001.

[23] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow" in Speech Transmission Lab. Quart. Prog. Status Rep., vol. 4, pp. 1-13, 1985.

[24] K.I. Nordstrom, G. Tzanetakis, and P.F. Driessen, "Transforming Perceived Vocal Effort and Breathiness Using Adaptive Pre-Emphasis Linear Prediction", in Audio, Speech, and Language Processing, IEEE Transactions, vol. 16(6), pp. 1087-1096, 2008.

[25] J. Laroche, Y. Stylianou, and E. Moulines, "HNS:speech modification based on a harmonic+noise model", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '93), pp. 550-553, Minneapolis, USA, April, 1993.

[26] T.K Moon, and W.C. Stirling, "Mathematical methods and algorithms for Signal Processing", Pap cdr Prentice Hall, 1999.

[27] C. Monzo, I. Iriondo, and Joan C. Socoro, "Voice Quality Modelling for Expressive Speech Synthesis", in The Scientific World Journal, Article ID 627189, vol. 2014, pp. 1-12, 2014.

[28] I. Iriondo, S. Planet, J.C. Socoro, E. Martinez, F. Alias, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification", in Speech Communication, vol. 51(1), pp. 744-758, 2009.

[29] O. Nieto, "Voice Transformations for Extreme Vocal Effects", M.Sc. Thesis Dissertation, Dpt. of Information & Communication Technologies, Universitat Pompeu Fabra, 2008.

[30] O. Turk, and M. Schroder, "A Comparison of Voice Conversion Methods for Transforming Voice Quality in Emotional Speech Synthesis", in Proceedings of Interspeech, pp. 2282-2285, Brisbane, Queensland, Australia, 2008.

[31] E. Moulines, and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", in Speech Communication, vol. 9, pp. 453-467, 1990.

[32] M. Schroder, and A. Hunecke, "MARY TTS participation in the Blizzard Challenge 2007", in Proceedings of Blizzard Challenge, Bonn, Germany, 2007.

[33] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech", in Proceedings of Acoustics, Speech, and Signal Processing (ICASSP), pp. 238-241Glasgow, UK, May, 1989.

[34] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA technique", in Speech Communication, Volume 11(2-3), pp. 175-187, 1992.

[35] E. Moulines, C. Hamon, and F. Charpentier, "High-quality prosodic modifications of speech using time-domain over-lap-add synthesis", in Twelfth GRETSI Colloquium, Juanles- Pins, France, 1989.

[36] J. Cabral, and L. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations", in Proceedings of the 9th European Conference on Speech Communication and Technology, pp. 1137-1140, Lisbon, Portugal, September 2005.

[37] X. Rodet, Y. Potard, and J. B. B. Barriere, "The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General," in Computer Music Journal, vol. 8(3), pp.15-31, 1984.

[38] J. Bonada, "Voice processing and synthesis by performance sampling and spectral models", Ph.D. Dissertation, Dpt. of Information and Communication Technologies, Universitat Pompeu Fabra, 2008.

[39] J. Laroche, "Frequency-domain techniques for high-quality voice modification," in Proceedings of 6th International Conference on Digital Audio Effects (DAFx-03), London, UK, 2003.

[40] E. Moulines, W. Verhelst, "Time-domain and frequency-domain techniques for prosodic modification of speech", In: W.B. Kleijn, and K.K. Paliwal,(Eds.), in Speech Coding and Synthesis. Elsevier, pp. 519-555, Netherlands, 1995.

[41] J. Laroche, and M. Dolson, "New phase-vocoder techniques for real-time pitch-shifting, chorusing, harmonising and other exotic audio modifications," in Journal Audio Engineering Society, vol. 47(11), pp. 928-936, 1999.

[42] J. Bonada, "High Quality Voice Transformations based on Modelling Radiated Voice Pulses in Frequency Domain", in Proceedings of the 7th International Conference on Digital Audio Effects (DAFx-04), Naples, Italy, October, 2004.

[43] R. Hoory, et al, "High Quality Sinusoidal Modelling of Wideband Speech for the Purposes of Speech Synthesis and Modification", in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, 2006.

[44] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: VOCODER revisited", in Proceedings of Acoustics, Speech, and Signal Processing (ICASSP), vol.2, pp.1303-1306, 1997.

[45] A. Robel, "A Shape-invariant phase vocoder for speech transformation", in Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz, Austria , September, 2010.

[46] P. Cook, “Spasm: a real-time vocal tract physical model editor/controller and singer: The companion software synthesis system”, in Computer Music Journal, vol. 17, pp. 30-44, 1993.

[47] Antares Vocal Processing > Products > THROAT Evo (2014, Jun. 19). [Online], Available at: http : //www.antarestech.com/products/detail.php?product=THROAT_Evo_14

[48] TC Helicon | Voice Pro - Ultimate Vocal Processing Tools (2014, Jun. 19), [Online]. Available at: http : //www.tc — helicon.com/products/voicepro/

[49] KaleiVoiceCope | Music Technology Group (2014, Jun. 19). [Online], Available at: http : //www.mtg.upf.edu/project/kaleivoicecope

[50] A.von dem Knesebeck, and U. Zolzer,“Comparison of pitch trackers for real-time guitar effects,” in Proceedings of the International Conference on Digital Audio Effects (DAFx-10), Graz, Austria , September 6-10, 2010.

[51] E. Gomez, and J. Bonada, “Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms Applied to A Cappella Singing”, in Computer Music Journal, vol. 37, no. 2, pp. 73-90, 2013.

[52] L. Mesbahi, V. Barreaud, and O. Boeffard, “GMM-Based Speech Transformation Systems under Data Reduction”, in Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.

[53] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice Conversion Using Partial Least Squares Regression”, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 912-921, July, 2010.

[54] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral Mapping Using Artificial Neural Networks for Voice Conversion”, inTransactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 954-964, July, 2010.

[55] L. Ming, W. Chen, and C.S. Chen, “The localised RBFs collocation methods for solving high dimensional PDEs”, in Journal of Engineering Analysis with Boundary Elements, vol. 37, no. 10, pp. 1300?1304, October 2013.

[56] M. Costa, P.Gay, D. Palmisano„ and E. Pasero, “A Neural Ensemble For Speech Recognition”, in Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS ’96), Connecting the World, 1996.

[57] MATLAB version 7.14.0.739 (R2012a 64-bit) . Natick, Massachusetts: The MathWorks Inc., 2012.

[58] L. Tamarit, M. Goudbeek, and K. Scherer, “Spectral Slope Measurements in Emotionally Expressive Speech”, in Proceedings of ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery, , Aalborg, June, 2008.

[59] V. Verfaille, U. Zolzer, and Daniel Arfib, “Adaptive Digital Audio Effects (A-DAFx): A New Class of Sound Transformations”, in in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1817 - 1831, September, 2006.

[60] U. Zolzer, “Pitch-based Digital Audio Effects”, in Proceedings of the 5th International Symposium Communications, Control and Signal Processing (ISCCSP), Rome, May, 2012.