Transformation of Whispering Voice to Pseudo-Real Voice for Unvoiced Telephony and Communication Aid for Voice-Handicapped Persons

March 2011

PASSOS Anderson Pierre

Abstract

For voice handicapped people, an easy to use voicing aid device is wanted. In mobile telephony, so-called non-speaking speech communication is an expected solution for essential privacy as well as for acoustic nuisance prevention. The study introduced here intends to cover both issues, introducing a system where the whispering (non-speaking voice or talk without vocal fold activation) signal is converted to pseudo-real voice signal, which is to be sent to, or heard by, the other party. The study also include validation test with multiple volunteers for its output legibility. Unlike general concept of speech regeneration being inclined to signal recognition or decomposition to text followed by electronic reading (voicing), our system converts it almost directly without recognition or decomposition steps. The processing is based on repetitive playback of short time autocorrelation, conducted by synthetic pitch pulse.

Acknowledgments

I wish to thank Professor Mutsumi Watanabe, Professor Kunihiko Mori and Professor Kazutomo Yonokuchi who served as examples inside the university and were always ready to offer me advice in difficult moments. Also, their constructive critiques made me eager to reach better results.

I also wish to thank my former advisor, Professor Yasuhito Takeuchi, for his unconditional guidance in this research during his time as my advisor in the Signal Processing Laboratory at Kagoshima University and also after his retirement. He taught me much more than Signal Processing and I am sure that his directives will guide me in the future.

I would like to thank my former lab matte: Miquel Espi. His help was invaluable and even after his graduation he continued to offer me his help and encouragement.

Last, but not the least, I wish to acknowledge the support received by the university staff: Mr. Junichi Matsuo, Mrs. Miho Yamamoto and Ms. Tomoko Imanishi.

INDEX

Abstract 2

Acknowledgments 3

List of Tables 6

List of Figures 7

1. Summary and Introduction 9

1.1 Summary 9

1.2 Introduction 10

1.3 Background 11

2. Whispering Voice 13

3. Telephony for the Handicapped 15

4. Telephony and Whispering Voice 17

5. Voice and Whisper side-by-side 18

6. Prior art by other parties 19

6.1 Phones for the Disabled 19

6.2 Indtal 20

6.3 Wearable Computing 21

6.4 FingerWhisper 22

6.5 Mime Speech Recognition 23

7. System Design and method 25

7.1 Identification of compromise band of frequency spectrum of whispering voiced signal 28

7.2 Whispering voiced signal capture 28

7.3 Pre-filtering of whispering voiced signal 30

7.4 Generation of artificial pitch 31

7.5 Dynamic autocorrelation 33
7.6 Reconstructed signal playback 36
7.7 Post-filtering the reconstructed signal 37
8. Experiment with volunteers 40
8.1 First Main Experiment 40
8.2 Second Main Experiment 41
9. Discussion and Results Validation 43
9.1 Validating Pitch Presence in the Reconstructed Signal 43
9.2 Analysis of First Experiment Results 44
9.3 Analysis of Second Experiment Results 48
10. Conclusion 52
Appendix I 54
Appendix II 55
References 57

List of Tables

Table 1 Set of words chosen for the first experiment 41

Table 2 Sentences chosen for the second experiment 42

Table 3 First experiment results summary 46

Table 4 Second experiment results summary 49

List of Figures

Fig. 1 Indtal Project showing tags to be accessed by voice commands (left) and a ruler simulating mouse events..... 20

Fig. 2 Nomadic Radio: Wearable audio messaging. 21

Fig. 3 The watch like device (1) captures the user's voice through a microphone (2) in the inner part of user's arm and uses bone conduction (3) to deliver the other party's voice to user's ears (4)..... 22

Fig. 4 Position of electrodes in the user's hand. Touching the month region with those electrodes make possible to detect month movement..... 23

Fig. 5 Proposed schematic of whisper to speech signal conversion system..... 26

Fig. 6 Frequency spectrum for whispering and normal voice for the sentence "This is a whispering test". The frequency spectrum of whispering voice does not contain the small horizontal lines present in the normal voice spectrum. Those lines characterize the pitch information..... 27

Fig. 7 A low-cut filter is applied to the lower part of the signal in order to clean up unnecessary/contaminated data. In (1) the original signal is shown while in (2) the lower components of the signal were cut-off. 30

Fig. 8 Quadratic pulse generated by our system in order to substitute the missing pitch information in the input whispering signal..... 32

Fig. 9 Whispered voice signal segmentation in frames..... 33

Fig. 10 Output for data “one” (A), “two” (B), “three” (C) and “four” (D) segmented (left) and with autocorrelation applied (right) 35

Fig. 11 Decision making point for the playback of output signal. Drived by the artificial pitch pulse, the system will decide among (1) play the processed frame, (2) play the original frame or (3) ignore the frame. 36

Fig. 12 Identification of unnecessary frequencies (1) in the output signal. 37

Fig. 13 Output playback(3) driven by the pitch pulse signal(2) artificially generated by the system. On every pitch pulse occurence,the system takes action and builds the output signal. 38

Fig. 14 Frequency spectrum of reconstructed voice signal (1) and normal voiced signal (2) for the same spoken/whispered sentence. Pitch pulse presence can be identified on both signals..... 44

Fig. 15 Butterworth filters of 2nd, 4th and 6th order. We can notice that a 2nd order filter is too smooth while the 6th order compromises the information in the signal 45

Fig. 16 Wave forms for data “one” (A), “two” (B), “three” (C) and “four” (D) before processing (left) and after processing (right)..... 47

Fig. 17 Whispering voice's spectrogram (left) and reconstructed voice's spectrogram (right) for the sentences: How are you doing..... 49

Fig. 18 Whispering voice's spectrogram (left) and reconstructed voice's spectrogram (right) for the sentences: Hello my friend (A), I am whispering (B), I really don’t understand (C) and Can I call you later? (D). 50

Fig. 19 Whispering voice's spectrogram (left) and reconstructed voice's spectrogram (right) for the sentences: I am very happy today (A) and Hello how are you doing? (B) 51

1. Summary and Introduction

1.1 Summary

This thesis mainly comprises the transformation of whispering voice into a pseudo real voice. For voice handicapped people, an easy to use voicing aid device is wanted. In mobile telephony, so called non-speaking speech communication is an expected solution for essential privacy as well as for acoustic nuisance prevention. The study introduced here intends to cover both issues, introducing a system where the whispering (non-speaking voice or talk without vocal fold activation) signal is converted to pseudo-real voice signal, which can be used in a wide range of applications, from telecommunications to voice aid devices.

The study also include validation test with multiple volunteers for its output legibility. Unlike general concept of speech regeneration being inclined to signal recognition or decomposition to text followed by electronic reading (voicing), our system converts it almost directly without recognition or decomposition steps. The processing is based on repetitive playback of short time autocorrelation, conducted by synthetic pitch pulse. At the end, the results of the system software developed by this study are presented and validated.

1.2 Introduction

In voice enhancement telephony, either for normal users, voice-handicapped or aged users, a possible first-aid model is text (or intermediate code) acquisition followed by text-to-speech conversion. This model would try to newly generate voice signal electronically using acquired text (or any intermediate code stream) rather than improve the quality of the input signal. The use of non-speaking (or whispering) voice is a novel expected method for speech communication and man-machine interface. Although yet an unpopulated area, some prior researchers challenged for text recognition for this sort of signal with some success. However, they didn't encounter the merit of regeneration of pseudo-real-voice based on this signal. The "recognition-reading" method is generally far from ideal due to its complexity and big computational load, and necessary database behind the processing chain.

For this purpose we conducted a trial to replace the missing vocal cord pulses in whispering action by injecting its substitute to conclude that it may be possible to use this different approach to non-speaking telephony since the generated impulse in association with the vocal cavity will give us a voice-like sound that can be better than artificial speech.

The study presented here converts the whispering voice signal to pseudo-real voice signal through a comprehensive deterministic process. The converted signal can be used to recover or re-create a pseudo-real-voice for voiceless mobile phone, speech handicapped patient or aged people.

1.3 Background

Whispering voice, also referred as unvoiced-speech, is a noise-driven response of vocal system in lieu of pitch pulse driven response. It is considered as understandable as normal speech when heard at proximity, and can also carry prosodic information. In voice enhancement telephony, including whispering telephony, a possible first aid model is text (or intermediate code) acquisition followed by text-to-speech conversion. This model would try to newly generate voice signal electronically using acquired text (or any intermediate code stream) rather than directly or indirectly deriving a voice-like signal from input signal. The use of non-speaking (or whispering) voice is a novel expected method for speech communication and man-machine interface. Some prior researchers [1] [2] challenged for text recognition for this sort of signal with some success. However, they didn't try to derive or to regenerate pseudo-real-voice based on this signal. The "recognition-reading" method is generally far from ideal due to its complexity, big computational load and necessary database behind the processing chain.

As it will be explained in the following section, prior researchers focused in acoustic-contaminated voice signal processing, but their processing was oriented into recognizing the voice carrying information and process it in a valid digital format (such as plain text). We believe that it is also possible to achieve our goal by using these existing tools in a speech to text and text to speech conversion but the main problem faced by this approach is its resource weight, what would generate a huge lag during processing time making it not suitable for real-time application.

Also, being our input and output signals both analogical ones, converting the analogical input signal into digital data for processing and then convert it back to analogical looks like an unnecessary step for us. This is one key point in this study: generating a process that, without taking into account the content or meaning of the input whispered signal is capable to output a pseudo-voice close to normal speech.

Another important point is that this study opens a large range of solutions for problems coming from disabled people. Vocal folds substitute devices are still an unpopulated area with precedents in studies for larynx substitution devices. However, the purpose of this study is to substitute the vocal folds keeping the reconstructed voiced signal as human as possible, what would enable people with chronicle diseases or irreversible injures on their voice organs, to create voice without use their vocal folds.

2. Whispering Voice

Whispering voice, also referred as unvoiced-speech, is a noise-driven response of vocal system in lieu of pitch pulse driven response. A very simple definition of whispering voice is "*a form of speech in which the vocal cords do not vibrate. It is usually not as loud as normal speech, but it does put stress on the vocal cords*". One of the main characteristics of whispering voice signal to be explored by this study is the fact that it does not contain pitch pulse information in it. The pitch pulse information and why it is important in the voiced signal will be explained in further sections

In using whispering voice signal, it has been known that some consideration for unnecessary or contaminating component in original microphone signal is necessary.

Takeuchi in 2003 [3] tried to protect proximity microphone from respiration air flow using smooth surface balloon (expanded condom) having no sharp edge or corner causing turbulence. It works fine, however, only when incorporated with sharp low-cut filter at microphone output to reject spherical resonance of the balloon in very low frequency, for example, 10 to 20Hz.

Cirillo in 2004 [4] also suggested, with own analysis, that contamination comes mainly from lower frequency part of ambient noise, speaker-own punctuation of vocal fold part of throat to mouth opening pathway, and aspiration/respiration air

flow jet noise, but not limited to them. Miquel in 2009 [5] showed also that the necessary part for reconstruction of a voice-like signal in the whispering voice is mainly, or even only, the formant parts.

Such characteristic of speech signal has been known since some time ago, in slightly different style. In amateur radio community, Harris and Gorski in 1977 [6] reported that 1500-2400Hz band is quite sufficient for their radio communication, other part like below 1500Hz can be taken out for bandwidth efficiency of the aerial signal, while fundamental part in 300-600Hz may optionally be attached in another modulation. These suggest that appropriate pre-conditioning of microphone signal would be a key factor in this business.

With the previous statements in mind, we developed our system in blocks which will be explained in Chapter 7 and its results discussed in Chapter 8 and Chapter 9.

3. Telephony for the Handicapped

Study of the whispering voice signal for recognition or control application is yet an unpopulated area. Kasuya [1] and Itakura [2] presented recognition issues for this sort of signal; however, they both did not cover the regeneration of pseudo-real voice based on this signal. When talking about telephony for handicapped patients or aged people, we know that this field is well known to have methods still relaying on a text-to-speech driven model or applying signal changes to "improve" the quality of the output signal.

Studies in this area, in large, still remain in recognition of text to drive text to speech conversion process. Unlike those studies, we are not going through recognition process; however, we try to convert the whispering voice signal to pseudo-real-voice signal through a comprehensive deterministic process.

In the past years a variety of assisting devices were proposed and partly marketed with much or less success targeting handicapped users. The business domain or like is supposed to have already decades of years of history but it still lacks on processing the voice signal itself, leaving voice-handicapped users with few options other than text driven applications or gadgets that supposedly would increase the usability of phones by those users.

Tentative of lip-reading where also studied by other parties but its development still seems slow and have no good way to detect consonant and nose tone. In Japan, NTT

Docomo developed a project [7] in which the facial muscle activity is captured by electromyography. The project has great recognition accuracy but it does not make use or even try to reconstruct the silent/handicapped patient's voice. Holzrichter [8] presented a solution where radar-like electric magnetic sensors are used to measure tissue motions in the vocal track during voiced speech. Some serious potential problems with this technique are electromagnetic exposure and, even more so, the fact that some articulatory states are very close to others and are exceedingly hard to discern even by direct observation (if possible). Instead of relying on existing techniques, we are trying to take another route to expand the possibilities. Takeuchi [9] tried to convert the whispering voice signal to recover/recreate a pseudo-real-voice for voiceless mobile phone and/or speech handicapped patient or aged people with some success and we believe the study presented here can help to improve his idea. Previously, we tried with some success to use low frequency signals as a substitute for the vocal folds [10].

Acceptable results can be obtained when converting whispering voice into normal speech if we process the signal directly. This approach opens a wide range of applications and further researches can be made to help people with chronicle diseases or irreversible injures on their voice organs

At this point, this research has a big opportunity. It would not only raise devices accessibility for people with physical disabilities but also spread the use of technology by providing important functionality for people with special needs.

4. Telephony and Whispering Voice

There are some interesting reasons making this research worth. Of course there are the commercial reasons, as in the cell phone market. It is well know that nowadays almost everybody uses the cell phone at least for job purposes, and in the most of the cases for personal interrelationships. It is well known also that the amount of places where any acoustic disturbance is absolutely prohibited is also huge (for example cinema theatres, concerts, the subway, taxi, hospital, and etcetera). The solution presented here can reasonably provide a way to interact thru the cell phone with other users in such way that the acoustic disturbance would be really little.

The ability to use whispering voice for phone would also allow the use of mobile devices in noisy environments and raise the precision of voice recognition applications.

5. Voice and Whisper side-by-side

The main physical feature of whispered speech is the absence of vocal cord vibration which in turn implies the absence of a fundamental pitch frequency and the consequent harmonic relationships derived from this [11]. This is the most significant acoustic characteristic of whisper, and also the most important characteristic in our study.

Regarding the voice signal, it is well known that the phase, even being an important character of sound, cannot be perceived by human's hearing system. The system defined in Chapter 7 will explore this characteristic of the human's hearing system to generate an almost real voice signal by playing back repeatedly the partial short time autocorrelation for this signal. This partial autocorrelation play back repetition in addition to additional processing to be explained in Chapter 7 will generate a valid sound signal (voice) with higher quality speech characterized by a more natural sound.

6. Prior art by other parties

In this section some studies conducted by other parties are shown. During the development of this study it was necessary to study about current and past researches as well as commercial products related to voice handicapped telephony and handicapped usability of current technologies. Some of those projects are listed here, as well as their good and bad points from the point of view of this research.

As a common point in the projects showed here, is the worry to make it easier to the user to use voice to interact with the tool. None of those projects works on the user's voice reconstruction or even with whispering voice in mind, and opens a broad range of possibilities for this research to improve and make handicapped users to that take full advantage of current technologies.

6.1 Phones for the Disabled

Since 1989, New York telephone companies are offering services for people who have difficulty using the telephone [12]. Some of them are:

(1) The relay operator receives a message typed by the deaf person on a computer screen, reads the message to the person on the other end of the line and then types the response back.

(2) Weak-Speech Handset: Increases the volume of the speaker's voice

(3) Impaired-Hearing Handset: Increases the volume of the voice on the other end of the line;

(4) For those who cannot hear the ring of a conventional telephone. A lamp plugged into this equipment flashes when the telephone rings

(5) Concentrates the sound energy of a ringing telephone into a frequency that can be heard by most people with impaired hearing.

6.2 Indtal

This project consists in a controlled tool for internet browsing targeting motor handicapped users having difficulties using a standard keyboard and mouse. The purpose is that those users can use voice to do everything. One interesting feature of this project is that users can "spell-out" words every time they need to type something. For general purposes a mouse controller is available through the use of commands that tell the application the screen coordinate the user wants the mouse to be placed.

Fig. 1 Indtal Project showing tags to be accessed by voice commands (left) and a ruler simulating mouse events

6.3 Wearable Computing

The "Nomadic Radio" project [13] was developed by the "Speech Interface Group" in the Machasussets Institute of Technology. This project consists in a wearable device that attempts to provide to the user highly personalized and timely information based on the context of user's tasks. User would be able to access voice-mail, news, appointments, weather information and other kinds of information through digitalized audio streams that would be downloaded from an audio server. The project uses technologies such as speech recognition for controlling and navigation and Text-to-speech synthesizer.

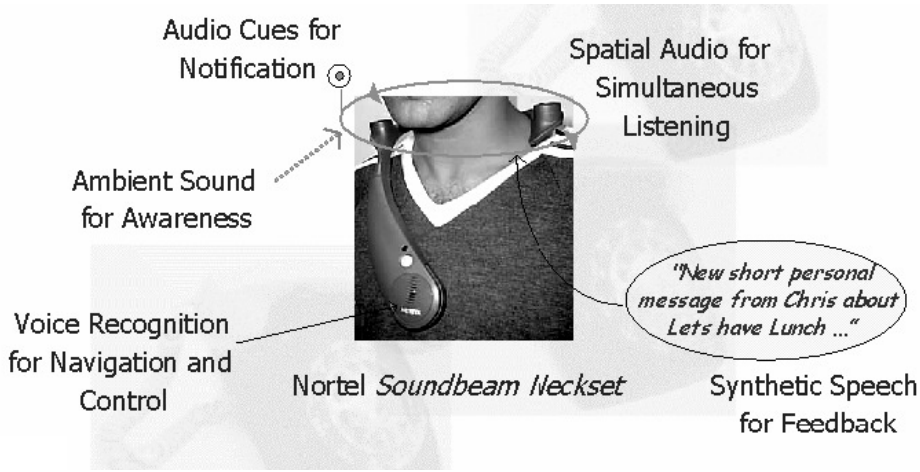


Fig. 2 Nomadic Radio[1]: Wearable audio messaging.

During experiments, researchers found out that the concentration of the user was reduced during other tasks execution. Another problem was the disruption to other people around the user because the message notifications and audio streams could

[1] Image used with author's permission

disturb people around. Privacy became an issue also, once the user wouldn't like people around to hear every received message.

6.4 FingerWhisper

The FingerWhisper project [14] in its quest for future communications possibilities is a new kind of wearable telephone handset that utilizes the human hand as part of the receiver. Since the microphone is located on the inner side of the wrist, the posture of the user's hand, when using the terminal, is the same as when using a cellular phone. This project uses bone conduction to replace a part of the receiver with the human hand.

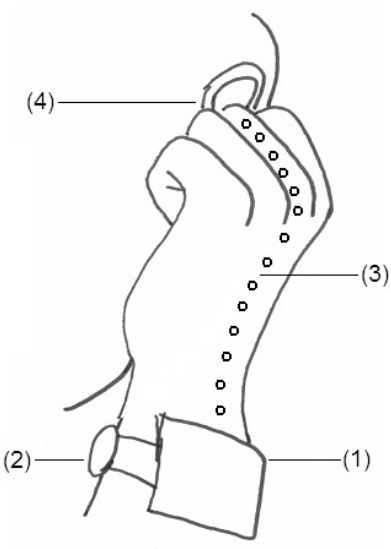


Fig. 3 The watch like device (1) captures the user's voice through a microphone (2) in the inner part of user's arm and uses bone conduction (3) to deliver the other party's voice to user's ears (4).

One of the big advantages of the FingerWhisper project is its watch-like shape makes it easy-to-wear and the mobile phone-like hand posture enables natural operation. Its watch-like design makes it easy to wear and frees the user's hands when not in use.

6.5 Mime Speech Recognition

The Mime Speech Recognition project [7] makes use of physiological information, such as facial muscle activity during speech. By using Mime Speech Recognition, electrical signals can be captured and recognized as voice even when users are simply miming speech. This project is also under development by NTT Docomo and currently, the five vowel sounds in Japanese language (a, i, u, e, o) can already be recognized by this technique.

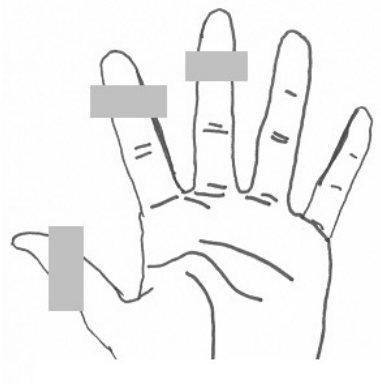


Fig. 4 Position of electrodes in the user's hand. Touching the month region with those electrodes make possible to detect month movement.

One characteristic of Mime Speech Recognition is the use of ring-shaped electrodes on the thumb, index finger and middle finger. Facial movement is detected by pressing the fingers against the face. When used in combination with conventional voice recognition, Mime Speech Recognition could greatly increase the recognition rate. Once perfected, such a system could be used in environments requiring silence, noisy environments, and as an auxiliary voice tool for the vocally challenged.

7. System Design and method

Our experiment bases itself in the assumption that the necessary part for voice reconstruction in the whispering signal is the formant part only. In order to cut out the unnecessary part we applied a generic low cut filter which acts as a pre-filter, preparing the signal for the core processing. It is also important to note that the pitch signal that is being used in this experiment is an artificial squared signal which simulates a pitch pulse signal. However, the possibility of obtaining such information from the input signal itself in analog mode, rather than digital mode, is real and is expected to be explored in further researches. Such method would enable the whole process to adapt dynamically to meet subject's pitch specifications, being able to recreate a much higher quality human-like voice from an unvoiced input.

The first step in our proposed method is to capture the whispered signal and filter it (Section 7.3) in order to clean the signal from contamination of airflow and ambient noise.

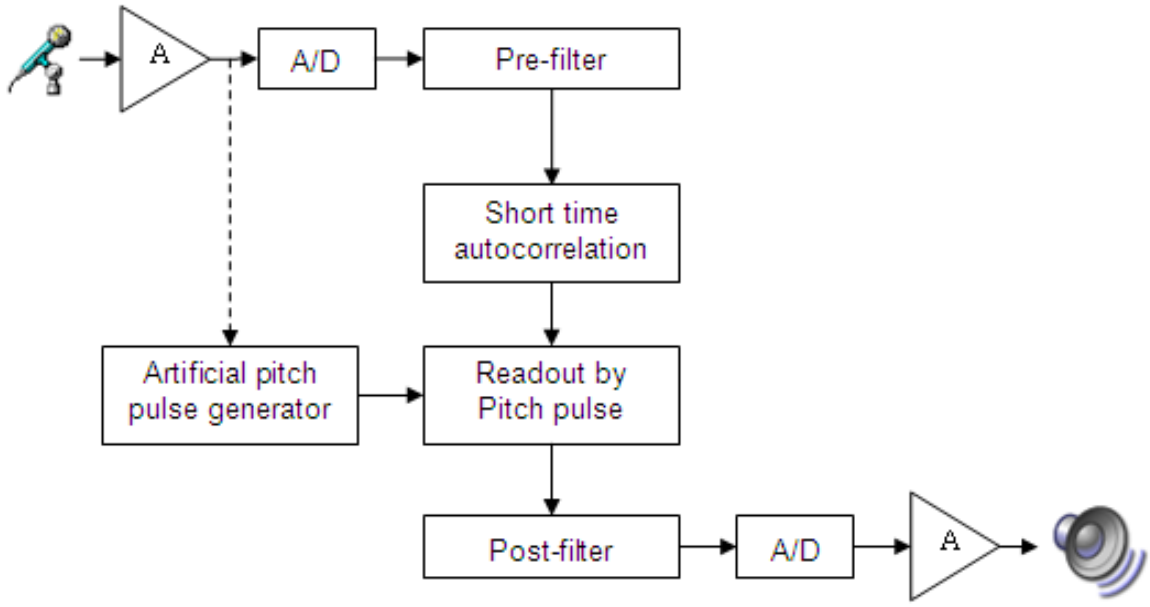


Fig. 5 Proposed schematic of whisper to speech signal conversion system

When analyzing the normal speech, one of its important characteristics is the pitch information carried by its signal. Comparing the spectrogram of normal speech with the spectrogram of a whispered signal (Fig. 6), we can notice the lack of stripes in the whispered signal's spectrogram. In a basic analysis, those stripes characterize the presence of pitch information in the signal.

Once the input whispered signal doesn't contain pith information one important step to be done during the processing of the whispered signal is to include the pitch information into, or fake as we opted in this study, the filtered signal. For that matter, an artificial pitch pulse generator had to be design in order to substitute this missing information in our input whispered signal (Section 7.4).

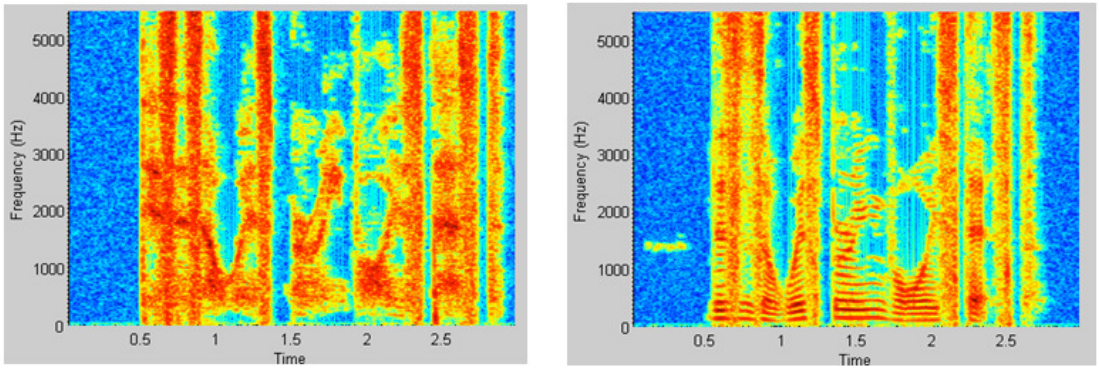


Fig. 6 Frequency spectrum for whispering and normal voice for the sentence "This is a whispering test". The frequency spectrum of whispering voice does not contain the small horizontal lines present in the normal voice spectrum. Those lines characterize the pitch information

During the processing, the input signal is segmented in frames and to each frame we apply an autocorrelation function (Section 7.5). Finally, at time of playback, the software makes a leveling evaluation of the signal's current frame and chooses the appropriated output (Section 7.6), generating a pseudo-voice like output similar to normal speech.

In order to process the whispering voice into a pseudo-voice similar to normal speech a software solution was developed. The solution makes use of MATLAB Signal Processing Toolbox. As input, samples of whispered and normal speech signals were captured with a conventional PC audio interface and a monaural microphone. The recorded words were sampled at 44.1 kHz. The whispered samples were them subjected to the processing above mentioned and as a result, we obtained

a pseudo-voice-like output similar to normal speech. To validate the output of the developed software, volunteers were hired to perform a live test and the results are presented in Chapter 9.

7.1 Identification of compromise band of frequency spectrum of whispering voiced signal

It is known that for voiced signals just a part of the frequency spectrum of the signal is necessary in order to maintain it understandable [15], but those frequencies are not applicable to the whispering voiced signals. Espi [5] has researched about the frequency spectrum which can be used to synthesize the pseudo voice. His study was focused not only on isolating the unvoiced speech signal, but to learn its properties and identify the frequency range where such signals can be found.

According to Espi [5] findings, the mandatory part of frequency spectrum necessary to process the whispering signal in is the range between 1500 Hz and 3000 Hz, but as it will be shown in Chapter 9, we found that a generic 4th order Butterworth low-cut filter at 1200 Hz is the optimal low-cut filter for our application.

7.2 Whispering voiced signal capture

The signal will be captured in its analog form and converted to digital. The input signal is expected to be the whispering voice signal originated by a person talking

close to the microphone. The microphone used by the system is a monaural microphone and its data is captured and stored in MATLAB®.

A pre-filter is used to pre-process the microphone signal in the context of isolating the formant frequency bands of the signal. A better microphone for this purpose is also considered and put to experiment, and in such way a study on how to improve microphone features on whispering voice signals recording can be also suitable.

The input block will also provide a low-weighted signal containing all the important information regarding to the acquired whispering voice signal input, and that is needed to the following procedures of the whispering voice signal processing. The whispering voice signal is formatted in form of a MATLAB matrix as input and will output a matrix containing a new signal with less carrying-information frequencies with the same size as the input.

The input signal is captured at 44.1 kHz and before applying any filter it is converted to 11.025 kHz. Since one of our main goals is to make this technology available in telephony, it is also important to notice that the resulting frequency is still covering the telephony standard, Nyquist (N/2). The output a signal will be equivalent to the input but band-pass filtered to remove unnecessary data or even contaminated data from the input signal.

7.3 Pre-filtering of whispering voiced signal

As stated before, the whispering signal suffers from several sources of contamination and the 1st stage of processing chain is to eliminate unnecessary component out from input signal. The pre-filtering process is important also because our system relies on the autocorrelation during its main processing and, having noise in the input signal give us unexpected results in the output.

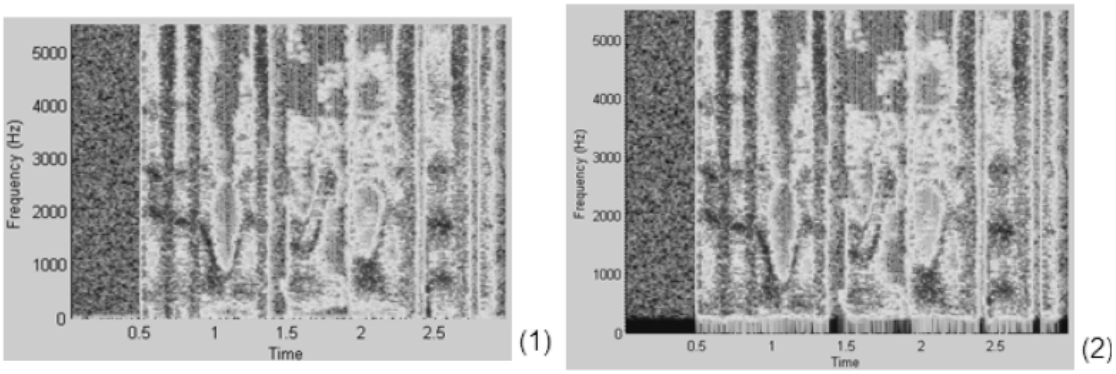


Fig. 7 A low-cut filter is applied to the lower part of the signal in order to clean up unnecessary/contaminated data. In (1) the original signal is shown while in (2) the lower components of the signal were cut-off.

The chosen filter for our system is a 4th order Butterworth low-cut filter at 1200 Hz. This value was achieved after making experiments with volunteers (Chapter 8) in order to identify the noise in the signal. Knowing that the necessary part for reconstruction of a voice-like signal in the whispering voice is composed mainly by the formant part, we felt confident to apply "human ears" and in order to clean up

the signal, we applied a generic low cut filter which acts as a pre-filter, preparing the signal for the core processing.

7.4 Generation of artificial pitch

Although vowels can still be recognized in whispering voice without pitch pulses, it is necessary for the listener to get information regarding the consonants to make the sentence understandable. There is some possibility to extract pitch-like in formation from frequency spectrum of whispering voice signal, by the structure or tension of the vocal cord and surrounding tissues as it may change according to hidden pitch by the speaker.

The pitch frequency information ("of intention") is specific for this kind of voice registered in F1 and F2 formants of the signal frequencies. There can be two reasons making whispered vowels understandable in that way and contributes to the free oscillation making it recognizable:

1. The structure of the vocal chord changes according to power, or
2. The level of tension of the vocal cord itself.

The fact is that there is no certain information about the vowel and this decision is left to the pitch pulse generator which, as said before, induces another extension for this research which we expect to be covered in future studies. A dynamically generated pitch pulse would be ideal, since the pitch changes according with many

factors such as word intonation. The artificial pitch information can be, for example, derived from instantaneous level of the signal [16], for it is likely to happen that when one raises the pitch frequency in intonation, one's voice also makes lauder. This behavior is quite well preserved even when one whispers without using vocal fold. So we can extract pitch substitute from signal level. This bold method works relatively fine for Japanese, and generates fair results for English also; however, pessimistic for Chinese because its accent and four kinds of tones (pitch trajectory), is another important information basically independent to loudness of voice.

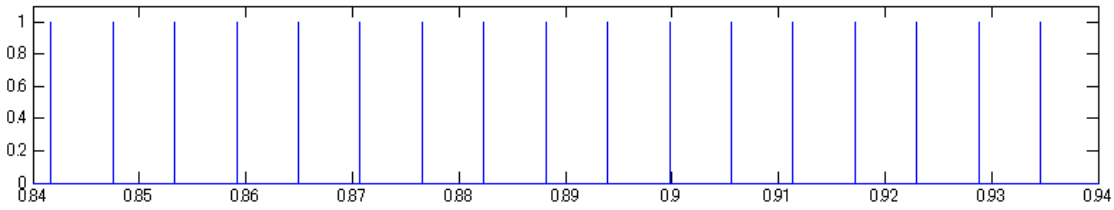


Fig. 8 Quadratic pulse generated by our system in order to substitute the missing pitch information in the input whispering signal

We believe that, to achieve a full reconstruction of whispered voiced signal into normal speech, an auto adaptive pitch pulse as described before is needed. However, here we do not engage this aspect in depth and try to extract much easy-going substitute to control playing-back pitch. Here, a quadratic pitch pulse (Fig. 8) is generated and used in the processing and reconstruction of the signal.

The designed pitch pulse generator can most simply be defined like a biased-VFC (voltage to frequency converter), although constructed in signal processing software.

It works independently from the input signal and during system execution does not change its behavior.

7.5 Dynamic autocorrelation

As our system uses dynamic or moving autocorrelation of the whispering voice signal, we make its Fano type decaying auto-correlation [17] as expressed below. This approach has better results in time domain resolution and computation efficiency, even compared with per-frame based power spectrum back conversion method.

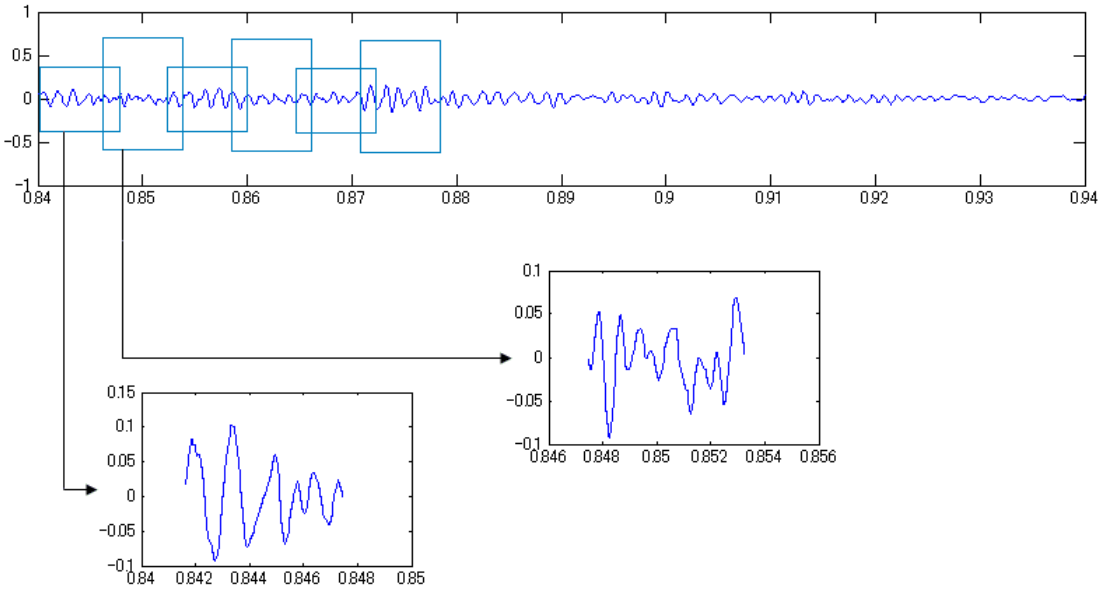


Fig. 9 Whispered voice signal segmentation in frames

To make it easier to process the signal in the developed software solution, the input whispered voice signal is segmented (Fig. 9) and put into an array. This approach happen to be very useful when reconstructing the output signal in Section 7.6 once the current array and its correlation counterpart can be easily accessed.

$$F(t_0,\tau)=\frac{1}{\alpha}\int_{t=t_0}^{\infty} exp\left\{-\frac{t}{\alpha}\right\}\times f(t)\times f(t-\tau)\times dt \quad (1)$$

F=Correlation function, *f*=source signal, *t*=progress in real time,

*t*0=present time, τ=time lag, α=constant

Here the continuously made correlation is always decaying while partly refreshed by incoming product terms. Its effective window size in time domain is determined by this decaying factor. In sampled, discrete expression, whenever a new sample comes in, the system operates the following procedure.

$$P(n,k)=D(n)\times D(n-k) \quad (2)$$

$$C(n+1,k)=C(n,k)+\frac{1}{2^N}\{P(n,k)-C(n,k)\} \quad (3)$$

Here α in (1) or *N* in (2,3) determines the decaying factor from present to past, and equivalent window size, and so-called self-forgetting time constant. For our prototyping, we used data size 256 samples while decaying time constant 64 samples equivalent. These numbers are subject for further optimization.

As we can see in Fig. 10, the autocorrelation of whispered samples generate a graphic with a central peak. This happens due to the lack of pitch information in the signal.

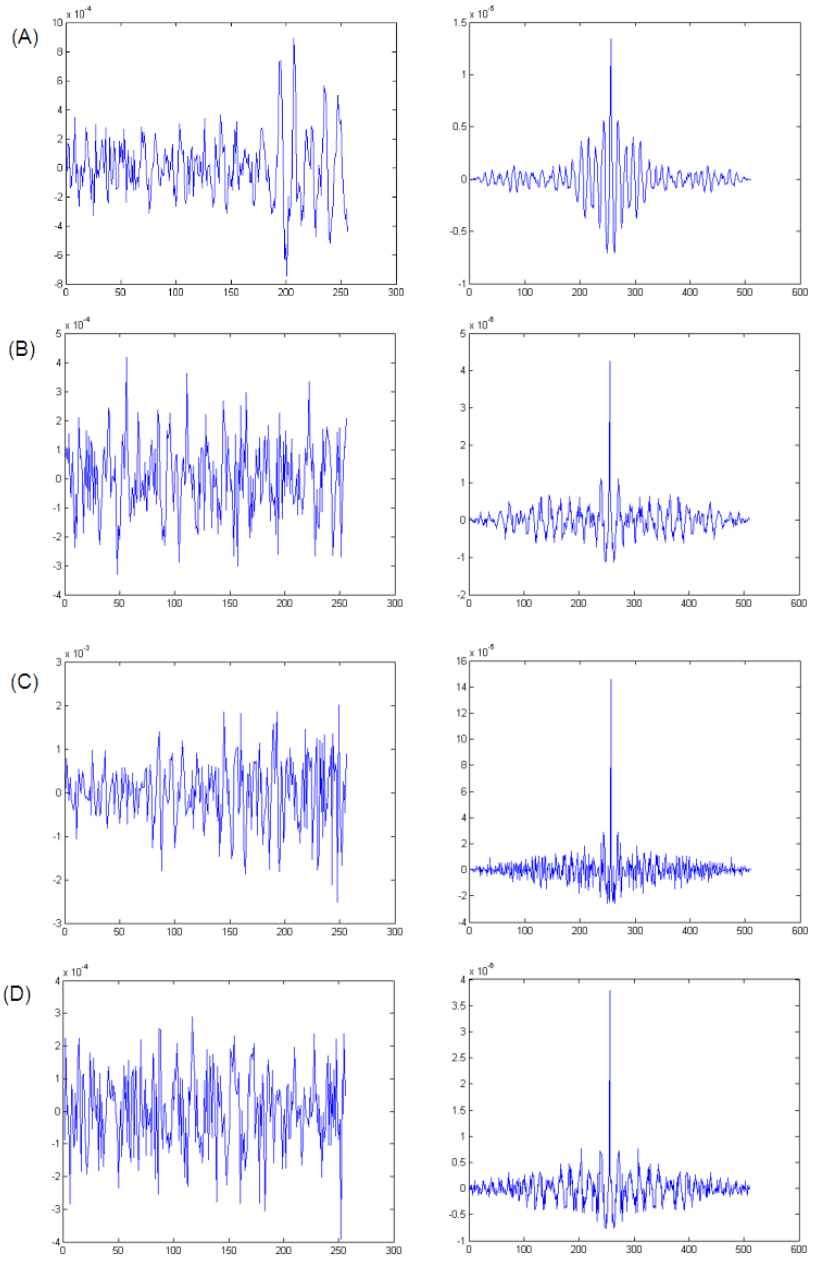


Fig. 10 Output for data "one" (A), "two" (B), "three" (C) and "four" (D) segmented (left) and with autocorrelation applied (right)

7.6 Reconstructed signal playback

In order to reconstruct the playback signal, our system makes an evaluation of the signal at the current frame and, based on its level, decides what signal to output. At this point, our system has the three signals running in parallel: the original input signal, the processed signal after the short correlation and the pitch-pulse generated artificially.

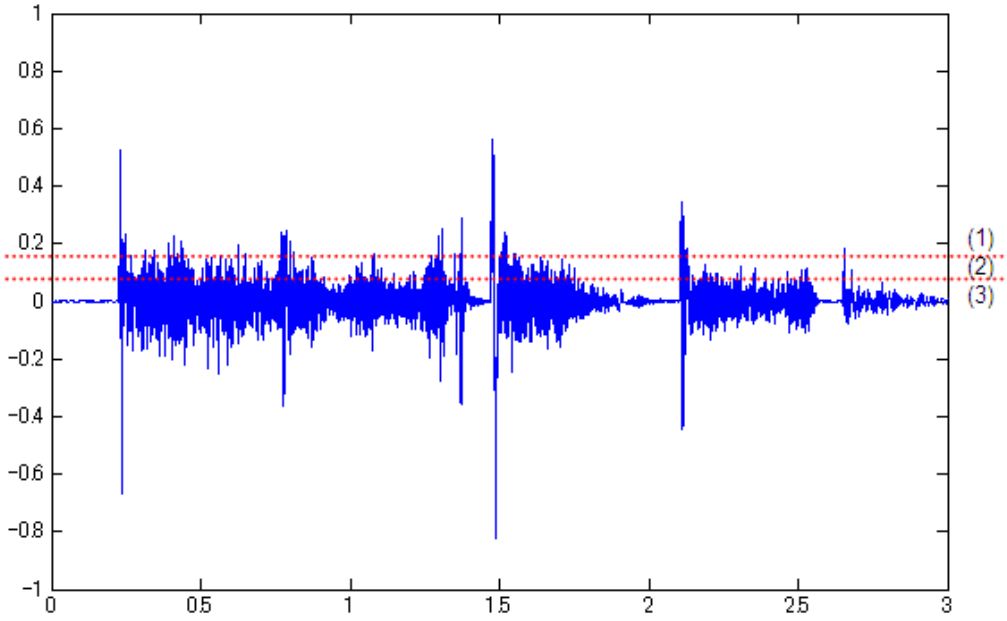


Fig. 11 Decision making point for the playback of output signal. Driven by the artificial pitch pulse, the system will decide among (1) play the processed frame, (2) play the original frame or (3) ignore the frame.

It is important to note that, instead of injection of the pitch-pulse signal, our system's playback is driven by the pitch-pulse. Basically, on every given pitch pulse

the system looks at the current frame, analyzes the level of the input signal and decides among three actions: Playback the autocorrelation of the frame, playback the frame as it is, or just to ignore the frame, as shown in Fig. 11.

This process is repeated cyclic until the end of the sampled whispered signal. As said before, the moving autocorrelation of the whispering voice signal give us a good response time avoiding any computational lag during playback.

7.7 Post-filtering the reconstructed signal

Since the correlation process is basically second order process, read-out signal for itself sounds more metallic and distorted than real voice. Here a compromising filter is hired (Fig. 12) to soften the reconstructed signal (Fig. 13) for more human-like voice. Its characteristic is, at this moment, determined simply by experience of hearing impression of the filtered output.

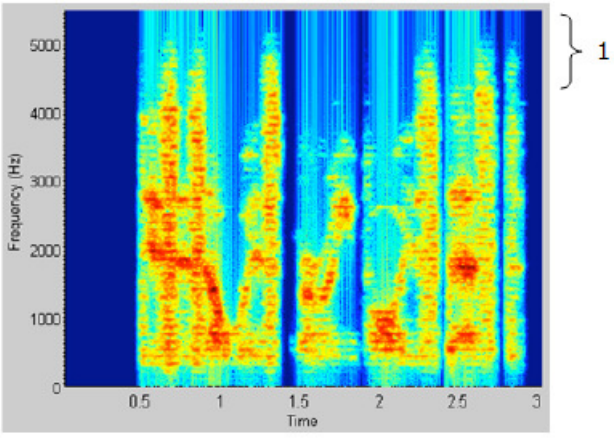


Fig. 12 Identification of unnecessary frequencies (1) in the output signal.

The importance of this block in the system is bigger than it looks. When talking about human-machine interaction, it is well known that the human being reacts much better to an interface similar to human than to a machine. In this case, the merit of recreating a voice signal from whispering would me shaded by the negative reaction of people having the strange feeling to be talking to a robot.

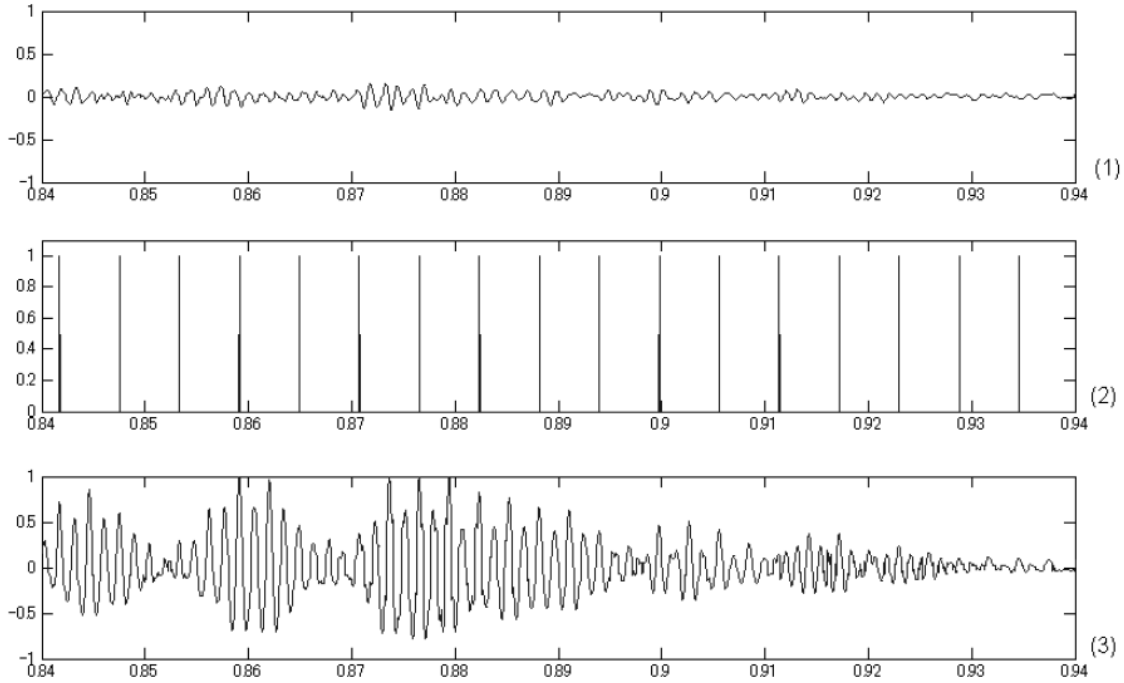


Fig. 13 Output playback (3) driven by the pitch pulse signal (2) artificially generated by the system. On every pitch pulse occurrence, the system takes action and builds the output signal.

A post filter block, with attenuation features and post processing of the recreated signal shows itself extremely necessary if we want this technology to be fully

adopted/accepted by people. This block cannot and should not be limited to this simple processing, and it is expected of future researches to address this issue deeper and improve the human-likeness of the signal through this block.

8. Experiment with volunteers

To validate the output of the method proposed by this study, two live tests were performed. The first experiment was conducted in our laboratory at Kagoshima University and had as its main objective, to determine the parameters for the filters to be implemented by our proposed system.

The second experiment was done in late 2010, and its main objective was to verify how well the volunteers' perception works when trying to recognize the sound samples. In this second experiment we also tried to evaluate qualitatively the data presented to the volunteers.

Following, both experiments are going to be explained in more details.

8.1 First Main Experiment

The first main experiment conducted in our study was done with 10 volunteers of different nationality, sex and age in order to create a heterogeneous data result. The set of sample words chosen for our testing (Table 1) was composed of twenty audio samples of words whispered into the microphone and not longer than three seconds. The set of words was randomly chosen and every volunteer was subjected to the full set at different frequencies with the words being played in random order to avoid memorization and invalidate the results.

Table 1 Set of words chosen for the first experiment

one	dog	amazing	born
two	cat	whispering	you
three	bird	sing	ten
four	lion	voice	boot
boat	bob	book	down

During the experiment the volunteers have been asked to evaluate the audio signal they were listening as “understandable” or “not understandable”. In order to find the optimal low-cut frequency to be applied in our filter, we made the low-cut filter frequency variable so it could be adjustable during the test sessions.

The results are summarized and shown in Chapter 9 and the questionnaire can be found in Appendix I

8.2 Second Main Experiment

The second main experiment conducted in our study was done with 10 volunteers of different nationality, sex and age in order to create a heterogeneous data result. The set of sample sentences chosen for our testing (Table 2) was composed of seven audio samples whispered into the microphone and not longer than three seconds.

After listening to each audio sample the volunteers were asked to say whenever they could understand or not what have been spoken and write down the words(s)

that they could understand. As a final step, the volunteers were asked to evaluate the quality of the sound they have just listened.

Table 2 Sentences chosen for the second experiment

How are you doing?	Can I call you later?
Hello my friend.	I am very happy today.
I am whispering.	Hello, How are you doing?
I really don't understand.	

The results for this experiment are also summarized and shown in Chapter 9 and the questionnaire can be found in Appendix II.

9. Discussion and Results Validation

Once the input whispered signal doesn't contain pith information and we know that it is an important component in the voiced signal, verify that our resulting signal contains such information becomes one important validation to be performed. Also, considering that the main purpose of this study is the recreation of the voice signal trying to keep it as human as possible, validate the output with volunteers (described in Chapter 8) to find the optimal parameters for the pre-processing and post-processing blocks sounded like an obvious and important validation to be done.

9.1 Validating Pitch Presence in the Reconstructed Signal

For the pith-pulse validation, a frequency-domain validation was performed through the generated spectrograms. Fig. 14 shows the frequency spectrum of the processed/reconstructed signal and the frequency spectrum of the normal voice signal for the same spoken sentence of same person.

As we can see in the frequency spectrum of the processed/reconstructed signal of Fig. 14 the resulting spectrogram now contains the small horizontal stripes that characterize the pitch information in the voiced signal. Such information was missing from the original signal (Fig. 6).

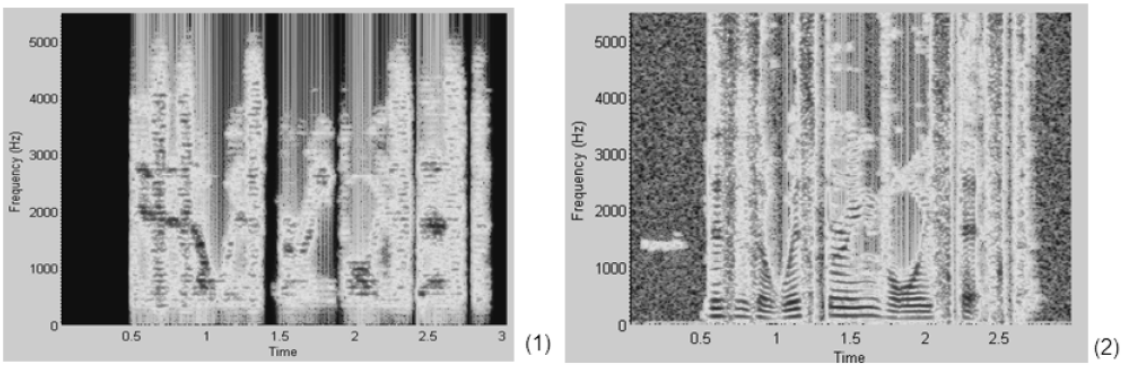


Fig. 14 Frequency spectrums of reconstructed voice signal (1) and normal voiced signal (2) for the same spoken/whispered sentence. Pitch pulse presence can be identified on both signals

9.2 Analysis of First Experiment Results

For the first experiment we hired ten (10) volunteers and a set of 20 English whispering words (Table 1). Those words where sampled at 44.1 kHz by conventional PC audio interface. The whispered word signals were put into a vector array in order to be processed by the MATLAB analysis toolbox. This approach allows us to apply the autocorrelation function to individual pieces of each signal (Fig. 9), and processes it accordingly.

In order to find the optimal parameters for the pre-processing and post-processing stages, the volunteers were inquired about the quality of the audio signal they were

listening and in a simple questionnaire sheet (Appendix I), they had to tell if it was audible (legible) or not and what was the spoken word being presented.

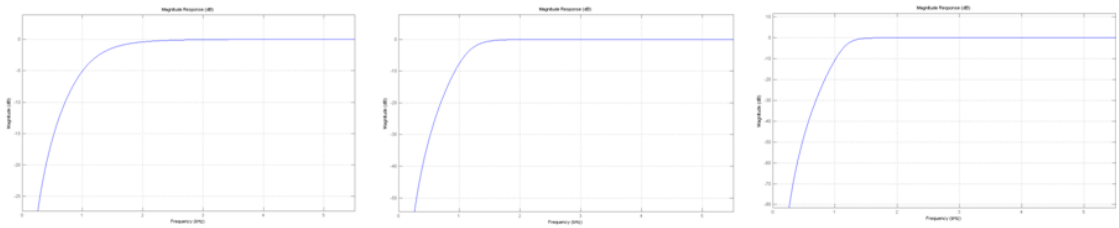


Fig. 15 Butterworth filters of 2nd, 4th and 6th order. We can notice that a 2nd order filter is too smooth while the 6th order compromises the information in the signal

During the questionnaire, parameters of the pre-processing and post-processing blocks were adjusted for best and balanced result. We reached to a 4th order low cut filter (Fig. 15) at 1.2 kHz for the pre-processing and a 4 kHz high cut filter for the post-processing.

Table 3 shows the questionnaire results and summarizes the answers provided by the volunteers. The numbers represent the number of volunteers who were able to understand the words at given frequency. For comparison, the wave signals before processing and after processing are showed in Fig. 16.

Table 3 First experiment results summary

	Understandable		
Word	**1000Hz**	**1200Hz**	**1400Hz**
one	100%	100%	70%
two	100%	100%	80%
three	100%	100%	80%
four	100%	100%	90%
boat	100%	100%	80%
dog	100%	100%	90%
cat	100%	100%	100%
bird	100%	100%	100%
lion	100%	100%	100%
bob	100%	100%	80%
amazing	100%	100%	80%
whispering	100%	100%	90%
sing	100%	100%	90%
voice	100%	100%	90%
book	100%	100%	100%
born	100%	100%	100%
you	100%	100%	70%
ten	100%	100%	100%
boot	100%	100%	100%
down	100%	100%	90%

As cited before, the frequency spectrum where the whispering signal is located starts at 1500 Hz. We noticed that, as the cut frequency gets close to that value, some volunteers could not evaluate clearly if the word could be heard or not. Lowering the filter to 1200 Hz makes it certain to have the entire whispering spectrum in our filtered signal.

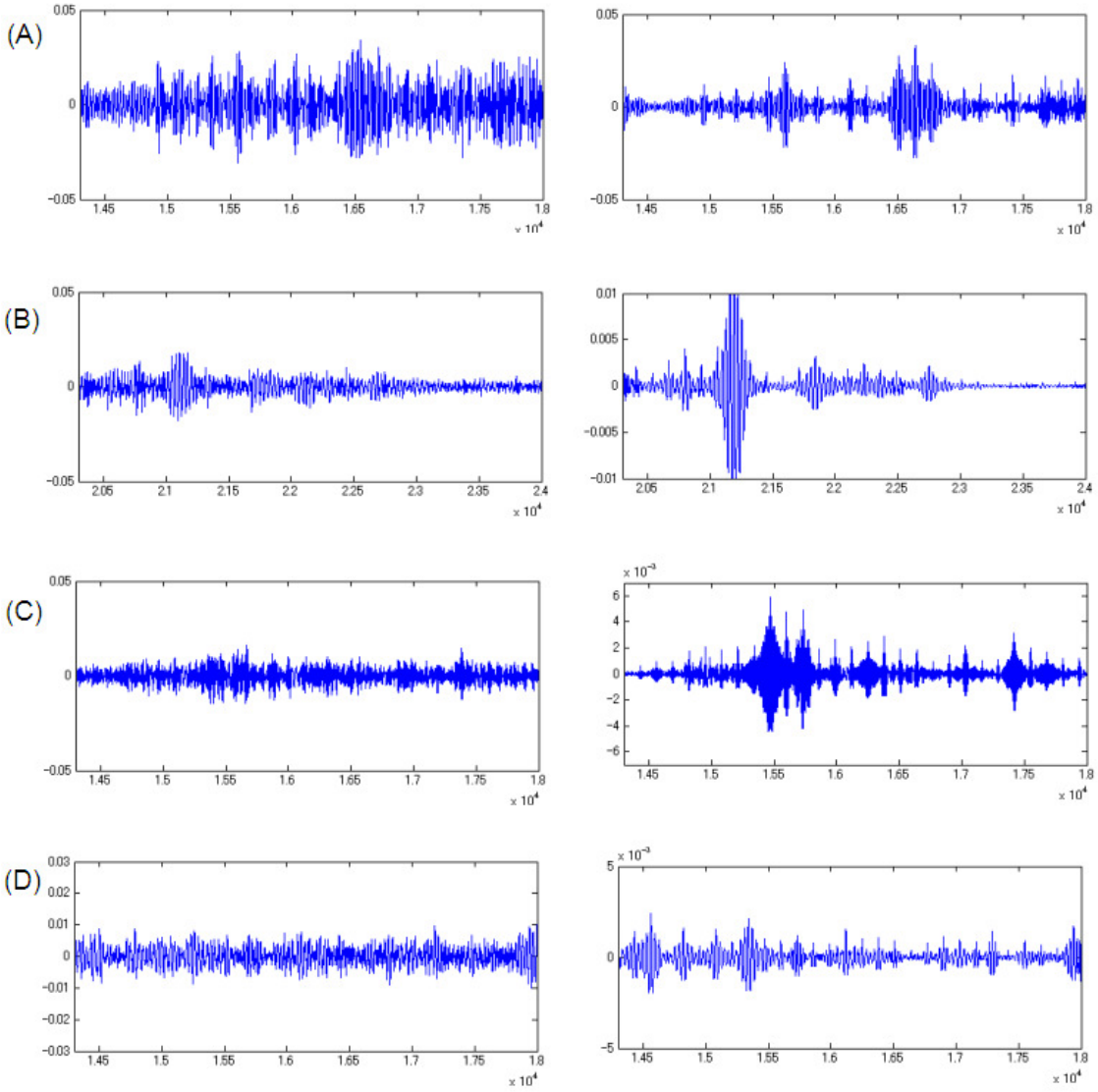


Fig. 16 Wave forms for data “one” (A), “two” (B), “three” (C) and “four” (D) before processing (left) and after processing (right)

9.3 Analysis of Second Experiment Results

For the second main experiment we hired ten (10) volunteers and used a set of 7 whispered English sentences (Table 2). The sentences where captured at 44.1 kHz by a conventional PC audio interface and later put into a vector that was processed by MATLAB analysis toolbox.

During this second experiment the audio playback parameters were not adjusted. The initial parameters were set based on the results of our previous experiment. All volunteers had to listen to the audio samples and check whenever they could understand or not the sound played. Later, they were asked to write down what they just listened (even if they thought it was not understandable) and as a final step, qualify the sound. For the last step, it was explained to all the volunteers that, when evaluating the quality of the signal heard, it was asked to the volunteers to keep in mind that what they were hearing was not real voice but a recreated voice sample from whispering voiced speech. A little background of our research was given to each volunteer before the experiment started.

Table 4 summarizes the answers given by the volunteers in this second experiment. Although most volunteers were able to understand and transcribe what has been spoken in the audio files, the acceptance rate is still low. Some volunteers stated that, although they could understand very clear what was being spoken, they thought the quality of the signal was too low. We attribute that to the fact that the recreated voice is still, in some way, too “robotic”.

Table 4 Second experiment results summary

	Understandable	Perfect Transcription	Quality						
			6	5	4	3	2	1	0
Can I call you later?	4	2					2	3	5
Hello, How are you doing?	9	9		1	1	2	3		3
Hello my friend	5	5				1	2	1	6
How are you doing?	10	9			1	2	2	3	2
I am very happy today	8	8		1	1	1	4	1	2
I am whispering	6	6				1	2	4	3
I really don't understand	5	5				1	2	2	5

The spectrogram of all the sentences used in this experiment can be seen in Fig. 18, Fig. 19 and Fig. 19. As expected, the whispering voice's spectrogram does not show the stripes (characteristic of pitch presence) while the reconstructed signal's spectrogram does.

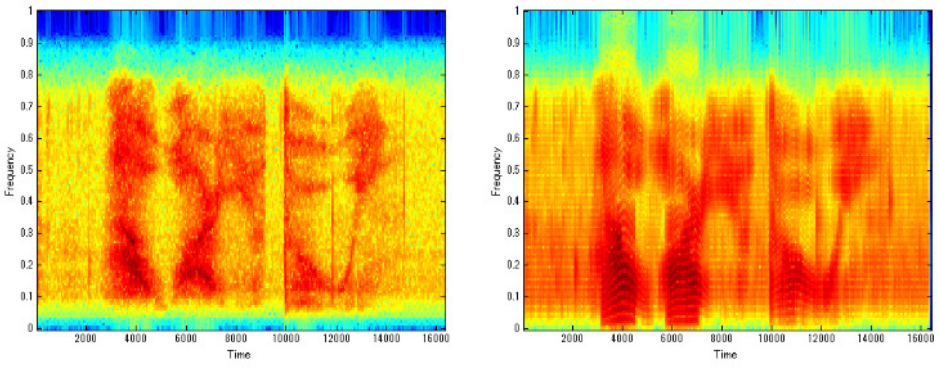


Fig. 17 Whispering voice's spectrogram (left) and reconstructed voice's spectrogram (right) for the sentences: How are you doing

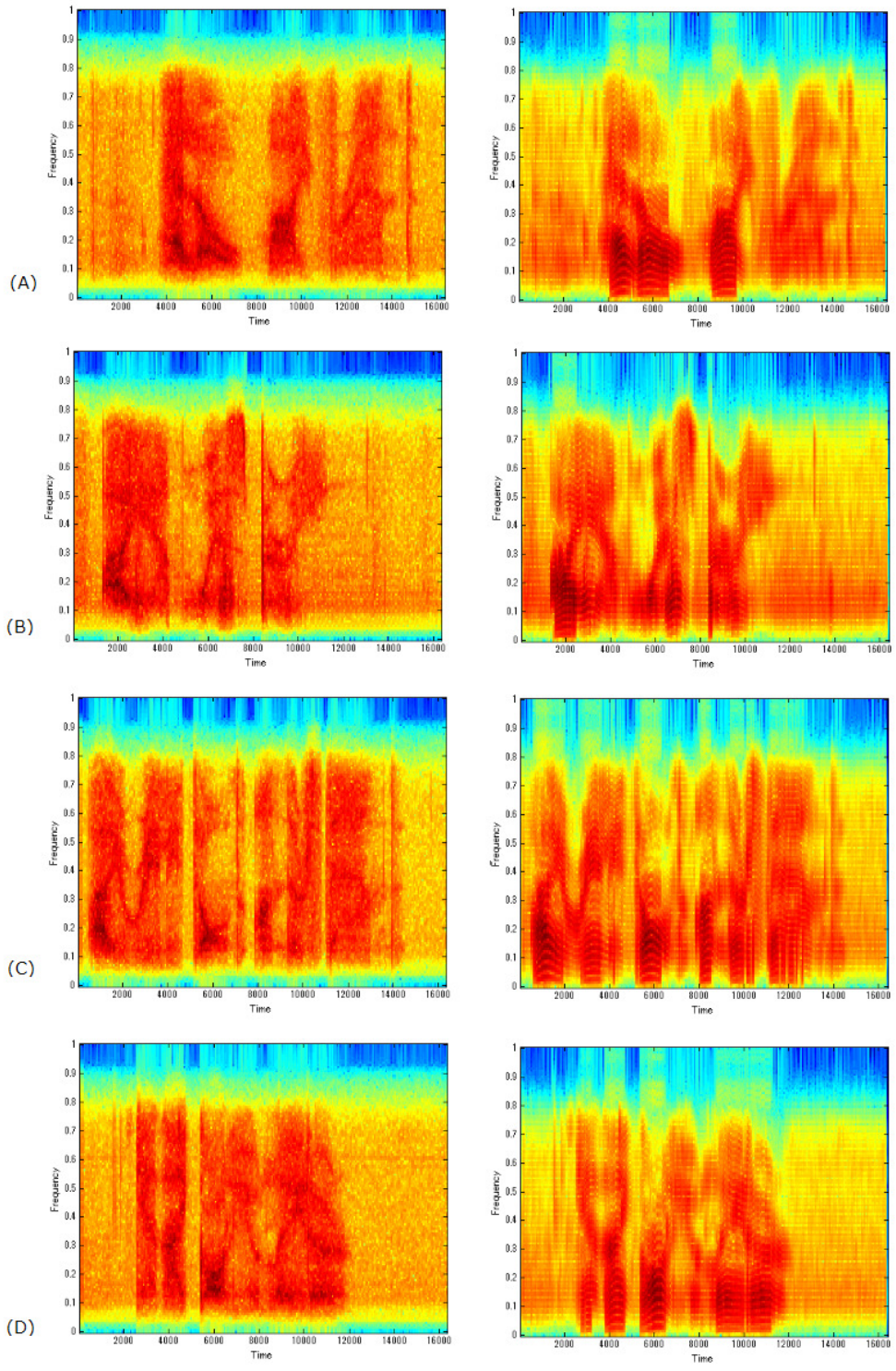


Fig. 18 Whispering voice's spectrogram (left) and reconstructed voice's spectrogram (right) for the sentences: Hello my friend (A), I am whispering (B), I really don't understand (C) and Can I call you later? (D).

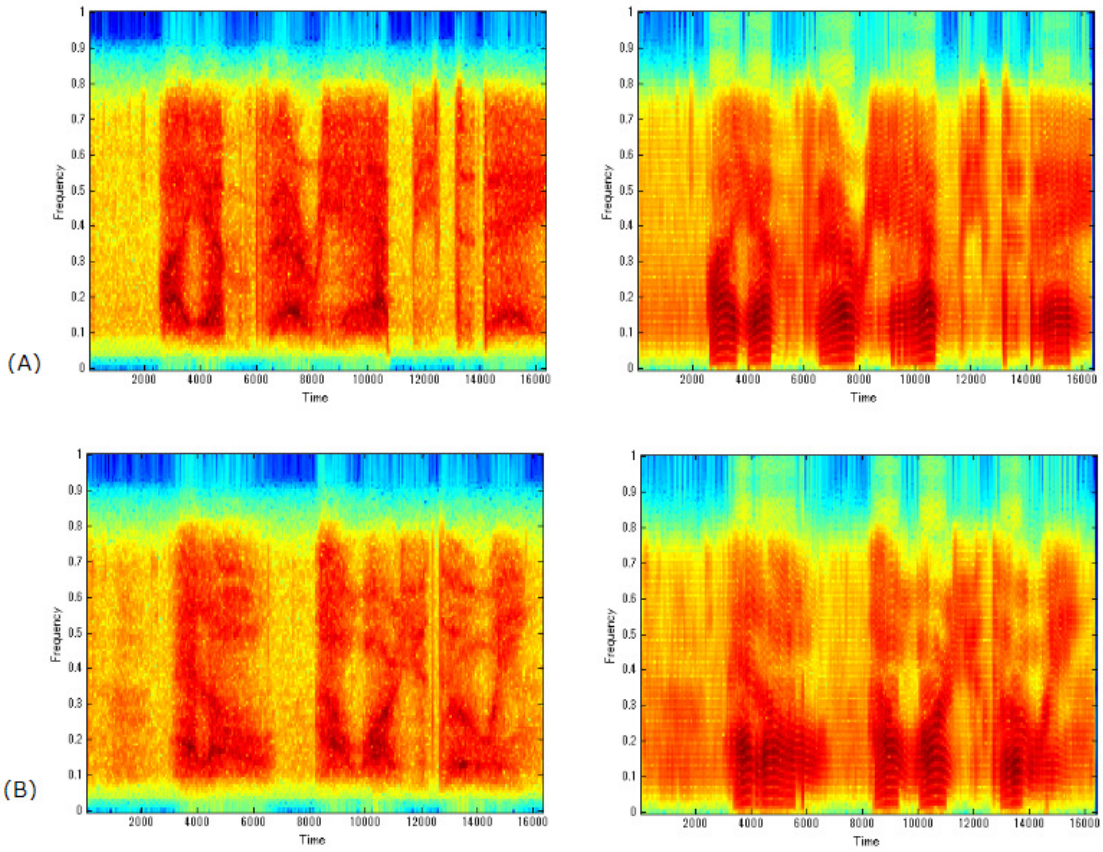


Fig. 19 Whispering voice's spectrogram (left) and reconstructed voice's spectrogram (right) for the sentences: I am very happy today (A) and Hello how are you doing? (B)

10. Conclusion

The method presented in this study, makes use of the instantaneous intensity of filtered signal to control the cyclic playback of output signal. Although we've used a quadratic artificial pitch-pulse as substitute to the signal's pitch-pulse, we believe that an enhanced artificial pitch signal is a key important factor to make the impression of human like voice. At this point the recreated pseudo-voice output by the system is still far from normal speech but further studies focused on the pitch creation are expected to improve the quality of recreated signal to more human-like.

Rejecting the lower frequencies in the input signal contributes a lot in the quality of the processed signal and the pre-processing filter showed itself a very important part during the pre-processing stage of the input signal. Contamination of the lower frequencies of the signal degenerate and compromise the results of the autocorrelation and without this step the results are very different from what expected.

Our second experiment showed us that, although the sound samples are understandable, the acceptance level by people is still low. Further experiments are necessary in order to attack this issue. We believe that in future studies the use of classical psychophysical methods are believed to give us more clues on how to evaluate and improve people's acceptance of the recreated signals.

For the handicapped, we believe our system shows an easy an inexpensive alternative for voice reconstruction. Future researchers are encouraged to find a better artificial pitch generator or (even better) a lightweight process for guessing the vowels during speech.

Although we may have a lot of room for improvement, the results obtained in this study are very promising. Being able to include the pitch information In the resulting signal was a great achievement and we believe that the method presented here is not only the starting point for other researches targeting whispering voice, but also it and can be improved and used in a near future to convert whispering voice signal into normal speech.

Appendix I

Signal Processing Laboratory
Experiment sheet

After listening to the audio sample, mark it as understandable or not understandable.

	Understandable	Not understandable
One	O	O
Two	O	O
Three	O	O
Four	O	O
Dog	O	O
Cat	O	O
Bird	O	O
Lion	O	O
Bob	O	O
Amazing	O	O
Whispering	O	O
Sing	O	O
Voice	O	O
Born	O	O
You	O	O
Ten	O	O
Boot	O	O
Down	O	O

Appendix II

Whispering Voice Recognition Experiment

Experiment Name: Recognition of pre-processed whispered sound samples.
Description: After listening to each of the seven sound files, the volunteer below identified will check whenever he/she could understand or not what have been spoken and write down the words(s) that he/she could understand. As a final step, the volunteer is asked to evaluate the quality of the sound he/she have just listened.

Name/Profession: _____

Date: _____

☐ I, above identified, participated in the experiment described in this document as a volunteer. I hereby allow the use of my answers in this experiment to be used for research purposes.

Sample Data 1

1. Can you understand what have been spoken? ☐ YES ☐ NO
2. Please transcribe what you've just heard

.....

3. Please, evaluate the quality of the sound you've just heard

Normal
Excellent Terrible

Sample Data 2

4. Can you understand what have been spoken? ☐ YES ☐ NO
5. Please transcribe what you've just heard

.....

6. Please, evaluate the quality of the sound you've just heard

Normal
Excellent Terrible

Sample Data 3

7. Can you understand what have been spoken? ☐ YES ☐ NO
8. Please transcribe the what you've just heard

.....

9. Please, evaluate quality of the sound you've just heard

Normal
Excellent Terrible

Sample Data 4

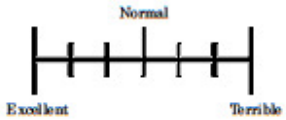
10. Can you understand what have been spoken?

☐ YES ☐ NO

11. Please transcribe what you've just heard

.....

12. Please, evaluate the quality of the sound you've just heard



Sample Data 5

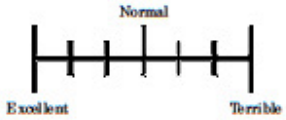
13. Can you understand what have been spoken?

☐ YES ☐ NO

14. Please transcribe what you've just heard

.....

15. Please, evaluate the quality of the sound you've just heard



Sample Data 6

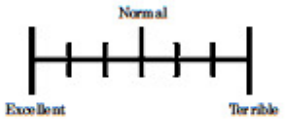
16. Can you understand what have been spoken?

☐ YES ☐ NO

17. Please transcribe what you've just heard

.....

18. Please, evaluate the quality of the sound you've just heard



Sample Data 7

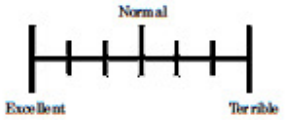
19. Can you understand what have been spoken?

☐ YES ☐ NO

20. Please transcribe what you've just heard

.....

21. Please, evaluate the quality of the sound you've just heard



References

[1] M. Matsuda, H. Kasuya, *Acoustic Property and Synthesis of Whispery Voice* , IEICE technical report, vol.99, no. 73, pp. 39-46, (1999).

[2] T. Ito, K. Takeda, F. Itakura, *Acoustic Analysis for Speech Recognition of Whispered Voice*, Acoustical Society of Japan 2001 Autumn Meeting, vol. 2001, no. 2, pp. 205-206, (2001).

[3] Y. Takeuchi, *Microphone Protected by Smooth Surface Sonadome to Pick-Up Whispering Voice (unvoiced voice) with Static Aspiration Pressure*, IEICE technical report, vol. 103, no. 331, pp. 19-22, (2003).

[4] J. Cirillo, *Communication by unvoiced speech: the role of whispering*, Annals of the Brazilian Academy of Sciences, vol.76, no. 2, pp. 413-423, (2004).

[5] M. Espi, Y. Takeuchi, *On the Mandatory Part of Frequency Spectrum of Whispering Signal in Order to Synthesize Pseudo-Voice*, IEICE technical report, vol. 109, no. 10, pp. 21-24, (2009).

[6] R. W. Harris and J. C. Gorski, *Narrow Band Voice Trans-missio*", QST, Dec. 1977, (1977).

[7] H. Manabe, A. Hiraiwa, T. Sugimura, *Unvoiced speech recognition using EMG - mime speech recognition*, CHI '03. ACM, pp.794-795 (2003).

[8] J.F.Holzrichter, *Speech articulator measurements using low power EM-wave sensors*, J. Acoust. Soc. Am. 103 (1) 622, (Jan,1998)

[9] Y. Takeuchi, *Voice regeneration system to convert whispering voice to pseudo-real voice*, IEICE Technical report, DSP2003-91, SP2003-86, September 2003.

[10] A. P. Passos, and Y. Takeuchi, *Signal Processing Specific to Whispering Voice*, IEICE Technical Report, vol.105, no.572 (SP2005 150-161); pp.13-17, 2006.

[11] V. C. Tartter, *What's in a whisper?*, Journal of the Acoustical Society of America, vol. 86, 1989, pp.1678-1683.

[12] C.Sims, *CONSUMER'SWORLD: Phones Become Easier For the Disabled to Use*, The New York Times, Dec.30 1989, p150 (1989).

[13] N. Sawhney and C. Schmandt, *Nomadic Radio: Scaleable and Contextual Notification for Wearable Audio Messaging*, CHI '99. ACM, pp. 96-103 (1999).

[14] M. Fukumoto, Y. Tonomura, *Whisper: a wristwatch style wearable handset*, CHI '99. ACM, pp. 112-119 (1999).

[15] Butler L., *Narrow Band Voice Transmission*, Amateur Radio January (1999-01), January 1999.

[16] P. B. Denes, and E. N. Pinson, *The Speech Chain*, W. H. Freeman Company, 2nd Edition, 1993.

[17] R. Fano, *Short Time Autocorrelation Function and Power Spectra* J.A.S.A., vol.22, No.5, pp546-550.

[18] A. P. Passos, *A Lightweight Processing for Conversion of Whispering Voice into Normal Speech*, ICALIP 2010, vol.1, pp74-79