Modeling of Whispering Effect

Titas Lasickas
Master student
tlasic16@student.aau.dk

Carmen Muñoz Lázaro
Master student
cmunoz19@student.aau.dk

1. ABSTRACT

Whispering has been studied with the generally aim of recognizing words in conversations of interest such as in the forensic field [1]. Also, whispered voice has been examined mostly to attempt the reconstruction of normal* speech [2–4], yet it has never been the other way around. In this project, the pursued goal is to get a whispered version from a normal voice recording.
∗ Normal voice/speech is considered speech voice when talking at a normal conversational level.

2. INTRODUCTION

Normal voice has never been used to build its whispered version hindering an accurate geared approach since the beginning. One approach could be made out of active filters trying to model the resonances of the vocal tract (which is varying rapidly when speaking). The determined algorithm to achieve whispering makes use of the identification of voice vowels formants in order to substitute them by its whispered version. Consonants are discarded for substitution as they mostly differ when whispered in volume rather than frequency content.

3. BACKGROUND AND LITERATURE REVIEW

3.1 Whispered voice

Whisper has lack of pitch (there are no vocal cords vibrations) and owns a turbulent excitation pattern that could be modelled by variant filters. This filtering has to resemble in real time the configurations (resonances) of the entire coupling of the trachea with the vocal tract due to the opening of the vocal folds. Consequently, due to missing pitches and spread energy, whispers have nearly flat spectra compared to normal voice [3]. Nevertheless, formants can be perceived in the spectrum as by hearing vowels and words can be recognized untroubled by our hearing system, meaning also that properties corresponding to tonal contrast in whispering have existence also in normal speech making possible its identification [5].

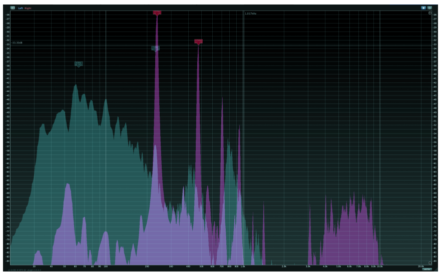

Figure 1. Spectrograms of “u” (pink) and whispered ”u” (blue)

3.2 From whisper to voice

As previously mentioned, little has been investigated related to the creation of whispering from voice. Even, it has been attempted to synthesize whispering from known whispers to achieve as a goal normal voice reconstruction. Studies usually look at first two formants, [6] spectral center of gravity, spectral energy spread and Mel frequency cepstral coefficients for the last aforementioned article [7].

3.3 Possible approaches

- Active filters
- General filters for selected parts of audio
- Substitution of identified vowels by its whispered audio version

To resemble the always changing vocal tract, active filters should be modeled for each time instant. Besides, filters should be personally designed as there are not two equal vocal configuration in the world. Moreover, turbulences with specific frequency content should be added. Complexity is highly increased by these challenging variables to replicate.
To make general filters first we need to find common features among vowels and differences between different vowels and consonants. Then, identifying only vowels (discarding consonants) and filtering them to make whispering effect.
Substituting vowels with prerecorded whispered vowels by hand would be the last resort, if no other method works.

4. EXPERIMENT DESIGN AND IMPLEMENTATION

Constructing active filters was rejected due to its complexity and a simpler, more approachable ways of creating whispering effect was chosen to be tested. In this section we will talk about our unique approach to create whispering effect.

4.1 Constructing study database

For the purpose of a better study, the distinct vowels were extracted, cut and classified individually manually while reading a Spanish text at normal voice level, resulting in 77 samples of each of the 5 vowels (a, e, i, o, u) at normal voice level.

4.2 Vowel recognition approaches

4.2.1 Formants

Formants are peaks in the spectral envelope that corresponds to resonances of the vocal tract. The auditory system uses them to identify sounds such as vowels. F0 corresponds to the main formant, F1 to the next found formant (dimmer than F0) and then consecutively F2 and so on. From the built database, formants (all that are detected by the software) are studied resulting in the Figure 2

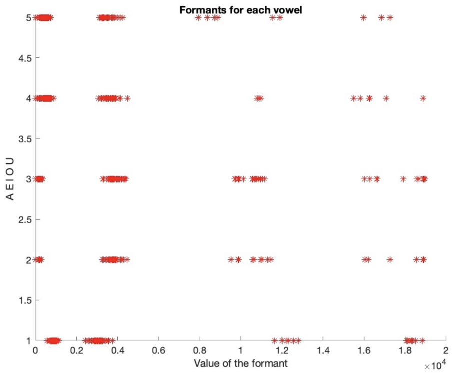

Figure 2. From bottom to top, AEIOU vowels and the frequencies they appear

The criterion used to identify a sound as a formant is that they belong to greater frequencies than 90 Hz and its bandwidth is less than 400 Hz [8].

4.2.2 Mel Spectogram

Based on the whole frequency content, different vowels are tried to be identified by the mel frequency cepstrum. MFC is a representation of the short-term power spectrum commonly used for speech recognition systems and music information retrieval (Figure 3).

4.2.3 Zero crossing rate

Zero crossing function by Jose R Zapata [9] was used to get zero crossing rate, which is giving us how often vector values cross zero in relation to total samples in a vector.
First we made zero crossing ratios for our database, for further reference and to check if there is as much overlap as it was with formants method described in section 4.2.1.
We use zero crossing rate function to identify vowels by windowing audio input to 200 samples sized windows and performing zero crossing rate function on them, saving where (in time) that value belongs. Figure 4 is showing our function on word "Alesi".

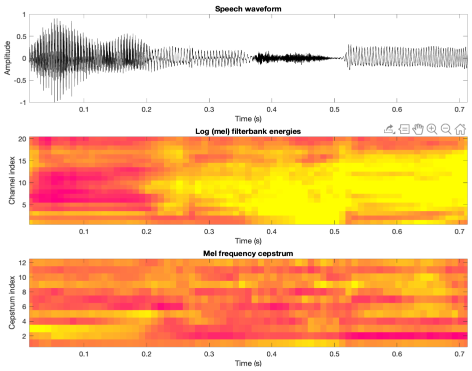

Figure 3. Log(mel) and Mel spectrogram of the word "ALESI"

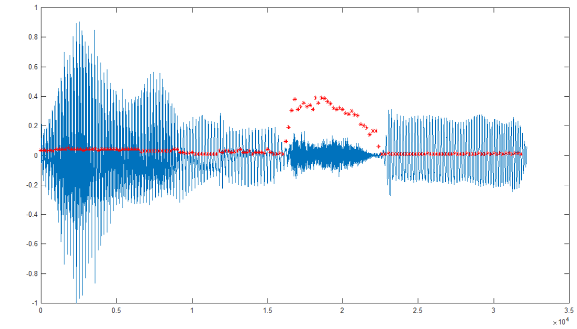

Figure 4. Word "Alesi" audio wave and zero crossing rate represented with stars

With zero crossing rate values on the whole database and slight adjustments based on windowed zero crossing on specific word we can set maximum zero crossing value for vowel and automatically separate vowels from consonants in the vector. This way we can filter only parts of audio, where vowels are at and leave parts with consonants unchanged.

5. RESULTS DISCUSSION AND FUTURE WORK

5.1 Formants

As we can see from Figure 2, vowel formant frequencies overlap, disabling its correct distinction by this feature. No different conclusion can be made neither when only regarding only F0 or F0/F1/F2. Here we mapped formants to a word "Alesi" in it different voice versions.

As we can see from formants graph from normal voice, they have nearly a constant value during vowels. However, they do not have a value enough distinguishable from others for proper vowel identification.

From last graph from whispered voice, formants are spreader as energy is flattened around spectra in whispering.

Reconstruction by hand of the whisper also shows this formants dissemination.

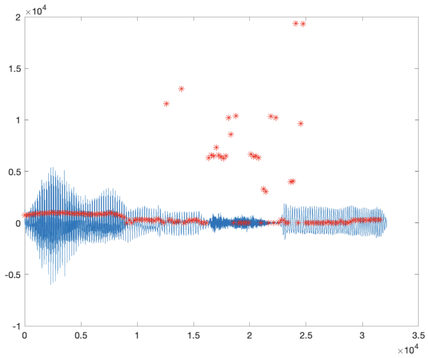

Figure 5. Formants found in “ALESI” word normal voice

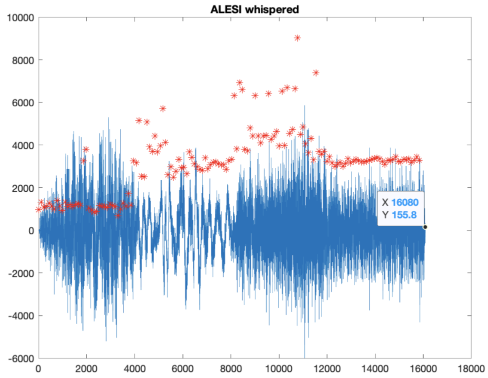

Figure 6. Formants in “ALESI” whispered

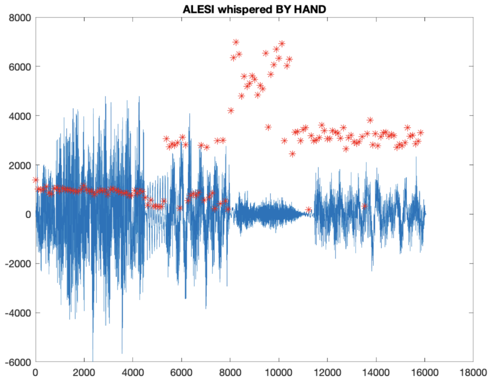

Figure 7. Formants in “ALESI” whispered constructed by hand

5.2 Mel Frequency Spectrum

When it comes to Mel frequency spectrum, regarding Fig. 8, it can be noticed that “A” have a certain amplitude at 5 cepstrum index. Also, slightly “I” has some relevance at 1 cepstrum index. “E” is more blurry in frequencies.

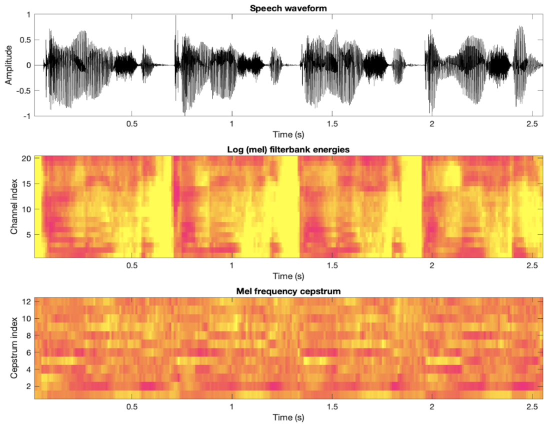

Figure 8. Audio wave, Log(mel) and Mel spectrogram of the words “ALESI, KALESI, TALESI, PALESI” spoken in time

There is some frequency content that can be useful for vowel distinguishing, however, a lot of statistical data must be collected in order to build a trustworthy database which would track all “versions” of a vowel (i.e: vowels pronounced at the end of a phrase or words tend to be in lower pitch, does not matter which class it belongs to) and therefore a reliable identification can be made.

5.3 Zero crossing rate

We have checked average zero crossing for our audio data set with vowel sounds. Looking at the Table 1 we can see that “E” and “I” have similar values, same is true with “U” and “O”, only “A” has value which is further away from other vowels. This means that when working with full words, vowels with similar values cannot be identified correctly.

A	0.0334
E	0.0101
I	0.0141
O	0.0275
U	0.0221

Table 1. Zero crossing rate average of all vowels individually

5.3.1 Zero crossing rate on word “Alesi”

We can just about see (Figure 4) difference between vowels “A” and “E”, consonant “S” is highly pronounced and later “I” has roughly the same value as “E”. This corresponds to our zero crossing values from data set in Table 1. vowels “O” and “U” should have similar effect, but we did not test that in a phrase.

Letter	Min	Max	Mean
A	0.025	0.05	0.0367
E	0.01	0.04	0.0187
S	0.06	0.39	0.2804
I	0.01	0.015	0.0112

Table 2. Zero crossing rate value statistics on windowed word ”Alesi”

Having the maximum value (in Table 2) of any vowel we can set it as vowel identification value and use it as described in section 4.2.3. This yields an audio wave in figure 9. When listening to the result we still do not perceive the sound as a whisper, rather a normal voice from further away with some metallic reflections.

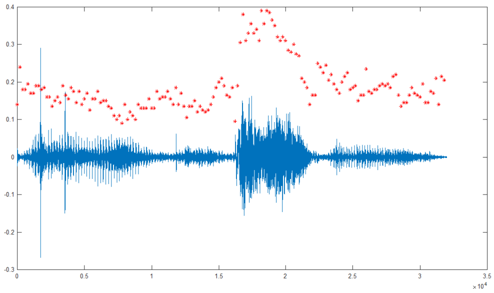

Figure 9. Word ”Alesi” automatically cut and filtered audio wave and zero crossing rate value represented with stars

Since we do not model lip smacks, higher amplitude of breathing and tongue movement noise, which are all audible in a whisper it is hard to recognise filtered sound as a whisper. All these sounds are not picked up by microphone at all so there is no way that we could filter and get them. This means, that these sounds would have to be added after filtering to better resemble a real whisper.

6. CONCLUSION

Studied methods for vowel identification could not be studied in the necessary depth to satisfactorily discriminate vowel classes and implement the proposed method. Manual replacement of vowels by its whispered version reach a credible audible construction of the whispering. However, the “transitions” among the replaced audio clips make it sound a little artificial. A replacement by the most similar version of a class vowel (pitch) could improve the realism for phrases resembling its full range tonality. A filter for “smoothing” transitions would also increase the realism of the whispered construction.

7. REFERENCES

[1] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, “Forensic speaker recognition.” Institute of Electrical and Electronics Engineers, 2009.

[2] T. Tran, S. Mariooryad, and C. Busso, “Audiovisual corpus to analyze whisper speech,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8101–8105.

[3] F. Ahmadi, I. V. McLoughlin, and H. R. Sharifzadeh, “Analysis-by-synthesis method for whisper-speech reconstruction,” in *APCCAS 2008-2008 IEEE Asia Pacific Conference on Circuits and Systems*. IEEE, 2008, pp. 1280–1283.

[4] J.-j. Li, I. V. McLoughlin, L.-R. Dai, and Z.-h. Ling, “Whisper-to-speech conversion using restricted boltzmann machine arrays,” *Electronics Letters*, vol. 50, no. 24, pp. 1781–1782, 2014.

[5] L. Jiao, Q. Ma, T. Wang, and Y. Xu, “Perceptual cues of whispered tones: Are they really special?” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] Y. Swerdlin, J. Smith, and J. Wolfe, “The effect of whisper and creak vocal mechanisms on vocal tract resonances,” *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2590–2598, 2010.

[7] S. Ghaffarzadegan, H. Bořil, and J. H. Hansen, “Generative modeling of pseudo-whisper for robust whispered speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, 2016.

[8] “Formant estimation with lpc coefficients.” MATLAB. [Online]. Available: https://es.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html

[9] J. R. Zapata, “Zero crossing rate - file exchange - matlab central,” https://se.mathworks.com/matlabcentral/fileexchange/31663-zero-crossing-rate, June 2011, (Accessed on 12/29/2019).