FCT Project POCI-01-0145-FEDER-029308

Dysphonic to Natural Voice Reconstruction

First-year report (2018-2019)

This report is organized on a per task basis.

A –Idiosyncratic voice signal analysis and modeling

– A.1 Dysphonic voice database and feature characterization

Leader: U.Aveiro

This task is devoted to the preparation of a database including both normal and dysphonic (i.e. whispered) voice recordings, and to the characterization of objective features that are required for signal classification and accurate parametric modelling.

Licenciate (Speech Therapy) Sara Isabel Pereira Castilho has been recruited from 1 October 2018 till 31 March 2019 and has implemented the following tasks (which are detailed in the attached report corresponding to file "Relatorio_SaraCastilho_DynaVoiceR_v2.pdf"):

-authorization has been requested to an independent Ethics Committee for the recording of voice samples of European Portuguese in both normal voice and whispered voice modalities and including an informed consent declaration to be signed by all volunteer participants,

-definition of the recording corpus to be included in the data base, planning of the recording sessions and preparation of the required technical conditions,

-motivation of volunteer speakers and realization of the recordings,

-manual phonetic segmentation and annotation (using the Praat platform) of all the recordings and generation of auxiliary files for all recordings such as to facilitate subsequent interaction with, and utilization by, other project tasks.

MSc. Researcher João Silva, a researcher in the project since January 2019, has developed a Matlab interface which permits to access all manually labelled recordings, to graphically represent the corresponding phonetic segmentation, and to add editing capabilities facilitating the interface with other software modules.

Outcomes:

M1_a: a database has been constructed of voice recordings pertaining to at least 15 different male speakers and 15 different female speakers, in both modalities (normal voiced speech and whispered speech), and including two-syllable works, short sentences and the "North wind and Sun" text. One report has been prepared by Sara Castilho. A paper is under preparation.

MM:

MSc. Researcher 1: 3 MM

Lic. Therapy Researcher 1: 4 MM

Susana Freitas: 1 MM

PI: 1 MM

Co-PI: 2 MM

– A.2 Accurate harmonic analysis and modelling of natural voiced sounds

Leader: FEUP

This task aims at creating an analysis and synthesis framework permitting full flexibility in controlling the frequency, magnitude and phase parameters of the harmonic structure that characterize voiced signals, as well as their evolution over time.

The baseline analysis-synthesis framework that already exists in this project allows to modify and control in a highly flexible way the harmonic and noise content of a voice signal. Thus, the activity in the context of this task has focused on the precise modelling and control of specific signal features with two objectives in mind: 1) to improve the accuracy of the analysis-modification-synthesis techniques and 2) to assess the perceptual impact of intentional modifications of meaningful signal features.

In this context, an MSc dissertation (Integrated Masters in Electrical and Computers Engineering) was developed by Francisca Vieira de Brito and successfully defended in July 2019 on "Precise harmonic modelling of human voiced sounds". The file of the dissertation is attached to this report and its name is "dissertacao_final_Francisca_Brito.pdf". The objective was to accurately model the microvariations of the fundamental frequency (F0) in a sustained vowel realization by one speaker, and to assess the perceptual impact when those microvariations are implanted in a realization of the same vowel by a different speaker of the same gender. Specifically, we wanted to assess if the microvariations carry a sound signature of a given speaker. In addition, we also wanted to assess if simple flattening of the F0 in a sustained vowel realization has a significant perceptual impact when compared to the original sound. Subjective tests have revealed that F0 flattening has a highly noticeable impact but the implantation of the F0 microvariations of one speaker into the voice signal of another speaker does not modify the sound signature in a noticeable way. This dissertation also confirmed that when the phase structure of the harmonics of the voice of one speaker are implanted in the voice harmonics of another speaker, a noticeable perceptual impact only exists in the case of male speakers. This confirms that only when the glottal cycle is long, the specific wave shape may have a significant perceptual impact for the same spectral magnitude, which indicates that the harmonic phase structure has idiosyncratic value especially in the case of low-pitched voices. Significant parts of this work were developed in collaboration with João Silva (MSc), a researcher in the project since January 2019.

Research has also been conducted trying to assess the relative importance in speaker identification of the spectral magnitude, spectral phase and F0-related information that is extracted in the sustained part of 5 vowel utterances. It is assumed that a phonetic-oriented segmentation scheme exists already. Results have confirmed that: 1) individually, spectral magnitude possesses higher discrimination capability than spectral phase or F0-related information, 2) fusing classification scores is more effective (86.5% correct identification has been reached) than fusing features in speaker modelling, 3) further research is needed to optimize score fusion. This has been reflected in the following publication:

“First Experiments on Speaker Identification Combining a New Shift-invariant Phase-related Feature (NRD), MFCCs and F0 Information”

Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018) - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS, pages 347-358

ISBN: 978-989-758-319-3

As a follow-up to the above publication, additional research focused on speaker identification using spectral magnitude, phase information and F0-related information, using independent classification according to each type of signal feature and then, finally, combining the independent classifications using score fusion. It was concluded that interesting results may be obtained as long as classification scores are congruent, which can be achieved with simple classification strategies that are implemented for each type of feature. These results have been presented in the following publication:

“Phonetic-oriented identification of twin speakers using 4-second vowel sounds and a combination of a shift-invariant phase feature (NRD), MFCCs and F0 information”

Audio Engineering Society International Conference on Audio Forensics

2019 June 18 – 20, Porto, Portugal

Outcomes:

M3_a: One MSc dissertation, two conference papers, harmonic analysis and synthesis algorithms, speaker identification computational procedures.

MM:

MSc. Researcher 1: 4 MM

Lic. Therapy Researcher 1: 2MM

Susana Freitas: 0.5 MM

PI: 2.0 MM.

– A.3 Accurate vocal tract filter modelling

Leader: U.Aveiro

The objective in this task is to implement an innovative ‘divide and conquer’ strategy allowing a refined estimation of the vocal tract filter using both harmonic and glottal noise information, this is very important for accurate modelling and projection of formant frequencies of linguistically equivalent voiced and whispered phonemes.

Marco António da Mota Oliveira (Licenciate), a researcher in the project since April 24, 2019, has performed a pre-processing of LPC-based spectral magnitude models of 9 oral vowels pertaining to a selection of 16 speakers, including both male and female speakers, in our database (that was obtained as a result of Task A.1). A database of spectral magnitude templates was built that corresponds to voiced vowels, whispered vowels, and that also corresponds to the residual signal of voiced vowels after the harmonic structure is subtracted from the signal. All these three types of templates were used in preliminary subjective experiments assessing the subjective quality of synthetic vowels using white noise as excitation and those templates as filters. Both sustained vowels and vowels in a word context were used.

This researcher has also initiated a cross-correlation study involving spectral magnitude templates characterizing different vowels and that are extracted from both sustained vowels and in-word vowels. This study is performed separately for each speaker. The purpose is to develop heuristics and an algorithm that is able to identify a specific vowel in running whispered speech such that corresponding voiced vowel templates can be fetched from the database and used in the synthesis of artificial voicing,

In addition to the above research, specific research in this task has also been conducted trying to assess the validity, in a physiological perspective, of the group delay response of all-pole vocal tract modelling. First insight on this was provided by the results of our DAFx2018 paper (which is reported in the context of Taks A.4) and suggested that the group delay response of all-pole modelling may introduce alterations in the phase structure of synthetic vowel sounds that exceed what is observed in natural voice signals. This was confirmed by the results of this specific research which suggested that LPC modelling is quite suitable to model the magnitude of the spectral structure of synthetic vowels, however, the phase structure should be controlled in an independent way such as to replicate the consistent phase structure that is observed with natural voiced signals. This research has been reflected in the following publication:

“On the physiological validity of the group delay response of all-pole vocal tract modelling”

145th Convention of the Audio Engineering Society

2018 October 17 – 20, NY

As a follow-up to the above paper, dedicated research has also been conducted in order to clarify what the perceptual impact is of using approximate models for the spectral magnitude or the phase structure, when natural voiced vowels are re-synthesized using those approximate models. In addition, frequency-domain and time-domain techniques were also compared. Subjective listening tests have revealed that approximate models for the spectral magnitude have a stronger (negative) impact than approximate NRD-based phase models. In addition, in the case of sustained vowels, higher quality signals were generated with frequency-domain techniques. Results have suggested that time-domain techniques need to be improved as they have the potential to perform better with real speech where syllables have a duration which is much shorter than that of sustained vowels. This research has been reflected in a paper submission:

“Subjective impact of holistic phase and magnitude descriptors in fully parametric harmonic speech representation and synthesis”

2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

Unfortunately, the paper was not accepted. It will be reformulated and submitted for another conference, possibly ICASSP2020.

Outcomes:

M2: First results on the correlation of LPC-based spectral magnitude templates describing different whispered vowels. One conference paper and one manuscript for another conference paper (it was submitted to Waspaa 2019 but was not accepted and is to be re-submitted to another conference)

MM:

MSc. Researcher 1: 2 MM

Lic Researcher (in lieu of MSc. Researcher 2): 3 MM

PI: 1.0 MM

Co-PI: 1.0 MM

– A.4 Accurate glottal source estimation and modelling

Leader: FMUP/Centro Hospitalar S. João

This task aims at characterizing accurately the true acoustic glottal excitation, especially the vocal folds vibration, thus under natural larynx-vocal tract interaction conditions; this sub- task relies critically on Medical (ORL) expertise as highly specialized medical procedures are involved.

Preliminary research has been conducted on the modelling of the phase structure of the harmonics of different sustained vowels that are uttered by the same speaker. Using a database of more than 30 speakers, it was possible to conclude that the phase structure is well modelled by the Normalized Relative Delay feature (NRD), that NRD possess a significant idiosyncratic value, that NRDs vary moderately for different vowels pertaining to the same speaker, that NRD appear to express essentially the contribution of the glottal pulse and that it is possible to extract an average NRD model of the typical human glottal pulse. This research has been reflected in the following publication:

“A holistic Glottal Phase-related Feature”

Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18), Aveiro, Portugal, September 4–8, 2018

Further research is needed to comply with the objectives set for this task and which require the purchase of a naso-fiberscope such that acoustic data near the vocal folds and under natural larynx-vocal tract coupling may be collected and analysed.

Outcomes:

M1_b: Acoustic data outside the mouth, models for the glottal phase, one conference paper.

MM:

PI: 1.0 MM.

– A.5 Adaptive phonetic segmentation techniques in dysphonic voice

Leader: FEUP

This task aims at developing computational rules enabling to select those regions in the whispered speech or dysphonic voice that should be subject to artificial voicing.

João Silva (MSc), a researcher in the project since January 2019, has been especially involved in this task and has been specifically involved with the development of a graphically-oriented research environment facilitating the research of features, heuristics and algorithms that can locate regions in a whispered signal that correspond to plosives. This is the first objective of a complex phonetic segmentation strategy the project team will develop in coming months.

Outcomes:

M3: First algorithms on plosives-oriented segmentation techniques.

MM:

MSc. Researcher 1: 2 MM

PI: 0.5 MM.

B –Perceptually natural synthesis of periodic voicing components

– B.1 Accurate vocal tract filter reconstruction

Leader: U.Aveiro

In this task we will combine and optimize the tract filter models emerging from sub-task A.3 such as to maximize linguistic correctness, speech naturalness, and speaker-consistent sound signature in the dysphonic voice implanted artificial voiced regions.

Activity in this task has not yet started.

– B.2 Perceptually natural voicing implantation and prosodic control

Leader: FEUP

This task focuses on the development of a method allowing to convert whispered speech or dysphonic voice, into artificially voiced speech; this conversion should be performed in such a way that the linguistic content of the speech is enhanced, the resulting speech sounds are more intelligible and natural than whispered speech or dysphonic voice, and that the speech conveys as much as possible acoustic cues allowing to easily recognize the speaker.

João Miguel Pinto Pereira da Silva (MSc), a researcher in the project since January 2019, started research work that is devoted to the accurate frequency estimation of sinusoidal frequencies that may exist in a signal. This research work was developed in collaboration with the PI and the MSc student Francisca Brito whose main work is described in the context of task A.2 of this project. Although whispered speech does not contain pitch estimation opportunities, good accurate frequency estimation algorithms are required in order to analyse normal voiced speech and to extract relevant models (e.g. concerning prosody) which are very important for high-quality natural voiced speech reconstruction using whispered speech as input. The state-of-the-art for non-iterative frequency estimation has been reviewed and an optimized algorithm has been developed whose performance exceeds that of close competitors in application scenarios that match those of our project. Innovative results have been obtained already and a journal paper will be written on this and submitted, possibly to the IEEE Transactions on Instrumentation and Measurement.

Research work has also initiated on robust and accurate fundamental frequency (or pitch) estimation by improving the algorithm the project team has been working with, and comparing it with other reference algorithms, namely Praat, YIN, and SWIPE. Substantial work is still needed before innovative and publishable results emerge.

Outcomes:

M5: First computational procedures paving the way to accurate and real-time fundamental frequency estimation and prosody modelling.

MM:

MSc. Researcher 1: 1 MM

PI: 1 MM

C –DyNaVoiceR system integration and real-time implementation

Leader: FEUP

This task is devoted to building a software app implementing the real-time reconstruction of natural voiced speech from whispered or dysphonic speech; this task will provide dysphonic voice patients with a material solution helping them to seamlessly use their voice to communicate effectively and comfortably in both human-to-human and human-to-machine scenarios.

Activity in this task has not yet started.

D –DyNaVoiceR usability tests and fine-tuning

Leader: U.Aveiro

This task articulates strongly with task C given that it will motivate early adopters to provide user feedback in order to fine-tune the DyNaVoiceR design and operation.

Activity in this task has not yet started.

E –Management

The management team consist of five members: the PI (FEUP), the Co-PI (U.Aveiro), a post-Doc (not yet recruited), Prof. Jorge Spratley (FMUP) and our distinguished consultant and eminent scientist, Prof. Paavo Alku from the Aalto University, Finland.

In addition to regular meetings that took place gathering project Researchers from different Institutions in the project (namely on 02 July 2018 and on 23 January 2019), three main facts deserve special consideration:

1-A project kick-off meeting took place on 18 June 2018 and this represented an important opportunity for all the project team to know each other (except Prof. Paavo Alku who could not attend), to review the project tasks and milestones, and to present a few very preliminary research results, the PowerPoint slides of this meeting are attached to this report, the file name is "DynaVoiceR_kickoff_meeting_18jun2018_v2.pdf",

2-With the important contribution of Marco Oliveira (Licenciate), a researcher in the project since April 24, 2019, a project web page has been initiated at the following web address: https://web.fe.up.pt/~voicestudies/dynavoicer/

Only the basic skeleton has been prepared, detailed information will be inserted into the website in coming months in order to make it informative and to reflect the project evolution and realizations,

3-Several new on the project appeared on the media, namely as a consequence of the 2019 World Voice Day (TV interview in April 2019 and another TV interview on 15 July 2019, in addition to several articles on newspapers).

In 2018, the PI has acted as Papers Co-Chair in the conference "DAFx 2018 - International Conference on Digital Audio Effects" which took place in Aveiro, Portugal, from 4-8 September 2018, and which addresses research areas closely linked to the DyNaVoiceR project.

In 2019, the PI has acted as Co-Chair of the conference "AES International Conference on Audio Forensics" which took place in Porto, Portugal, from 18-20 June 2019, and which addresses research areas closely linked to the DyNaVoiceR project.

The annual project meeting for 2019 will be scheduled in coming months.

MM:

PI: 0.4 MM

Co-PI: 0.15 MM