Preliminary processing of the human voice recordings

Damian Krzesimowski, Kielce University of Technology

Abstract
Analysis of the human voice is a difficult process. This is due to the complexity of the voice signal, and its individuality against every human being. Keep in mind that a person can freely change the tone of voice, speed of speaking and many other parameters that are independent of content of speech. The first task of the researcher is to eliminate differences in the voice signals for all recorded people. This is the only way you can get a sample of data that can be compared within a pool of recordings performed in the same way. This paper describes a proposal for the selection of audio data for a short speech on the basis of one voiced sound. The study was conducted a directional microphone and recorder having the ability to save as uncompressed audio. Recordings were carried out 31 people, and then the Fourier analysis, spectrogram and cepstrogram of selected data were determined. These characteristics were used to develop a path for dealing with voice recordings of differing length, amplitude level and a way of expression of the recorded persons. The study was based on comparison of the graphical results for various parameters of selection. The result is a strictly specified path of conduct, which can be used for the short-term analysis for most sound recordings conducted under different environmental parameters.

1. Introduction
The study of voice is an activity extremely difficult, and this is due to instability of the human voice. The same person denouncing the same issue within a few minutes otherwise modulates some sound, says the issue of varying intensity of sound. The problem of the study of voice simplifying when dealing with only one sound. In this case, the voice recording can be divided into individual phonemes, you can also record only the desired phone. Due to avoid the need to identify the sounds the author opted for the latter. The aim of this study is to prepare the recordings to eliminate as much as possible errors resulting from distortion, sound samples should also be standardised in terms of duration and amplitude. During the study used a set consisting of the sound recorder and directional microphone were used. The recordings were made at a frequency of 96 kHz and 24-bit quantisation of the amplitude. Practical application of methods of inference from voice recordings requires assumptions about the performance of sound recordings on which the information input in a controlled, uniform characteristics of ambient noise and interference introduced by the recording equipment. Changing these parameters should not require intervention, provided all appropriate preventive measures were used. It consist in providing high quality recordings, the selection of appropriate sensitivity factor of the recording process and the using covers or directional linear microphone in order to effectively reduce noise in the recording. Recorded voice in the digital representation is subjected to preliminary processing, involving the conversion of the input signal to the mono and removing a part of record for analysis along with normalisation of the maximum amplitude of the waveform.

2. Quantisation of analogue signal
The first stage of the system is recording the audio signal in digital form. This is done using an electronic recorder with a directional microphone with linear frequency response. A decision on the discretisation of the speech signal results from less interference to the signals in digital form and ease of analysis of the data stored in digital form. The processing of continuous signals to discrete consists of three operations: sampling time, quantisation of values, and coding. The operation of signal sampling involves taking of fragments of the continuous signal $x(t)$ in predetermined equidistant moments. Speech signal sampling time was set at 10.47 microseconds, that is, the sampling frequency is 96 kHz. The quantisation is bringing a set of values of the signal $x(t)$ to its finite subset. It produces a signal $x_q(t)$, i.e. the signal with values from a given set of numerical values. Signal quantisation results from the need for analogue to digital converter prior to analysis using digital arithmetic logic unit. Quantisation is used in 24 bits per sample; it is therefore possible to obtain a set of finite 16,777,216 in real terms. The resulting digital data are stored in the fixed-point format to a matrix form.
During the digitising process should take into account the loss of some part of the information about the signal. The sampling theorem shows that if the signal frequency bandwidth is limited, it is enough sample rate of at least two times greater than the width of this band. A man hears sounds from 20 Hz to 20 kHz. Frequency band generated by humans is less, due to the signal strength can be assumed width of 6000 Hz. The greatest power to acoustic signals is in the frequency range 60 Hz to 540 Hz, and the level of signals at higher frequencies decreases dramatically. This has considerable impact on the speed of analysis and level of errors occurring outside the range of frequencies generated by a human. In accordance with that the adopted sampling frequency 96 kHz allows to obtain samples with sufficient high precision. It should be noted that the studies used recording equipment with 128-times oversampler. It allows to obtain 12,288,000 samples per second. Thus, the quantisation accuracy is greater than the quantisation without performing oversampling. You can calculate the quantisation error of the signal value. In Figure 2 the maximum amplitude of the signal level is 0.23 over and –0.43 below zero. In a 24-bit value a minimum stroke is 0.0000000268. This value is so small that it is possible to assume that the desired accuracy is sufficient.

3. Channel selection of the desired data

Registered the signal is recorded on two audio channels, each in a separate column of the matrix data. Time course of recorded speech sounds "a" and the course of an unused channel is presented in Figure 1. It is the smallest possible number of audio tracks, while recorded, offered by the hardware. Multichannel recording is superfluous given the analysis in many studies only one audio track The conversion paths to a single channel in a way that the hardware comes down to remove every second sample of the two paths and enter the other, alternately, to preserve the sampling target. This method is disadvantageous for the present studies because of the loss of valuable information.

It should be noted that only one of the channels contains useful numerical data. The second channel, although it is not used, has an amplitude distortion of less than three rows of the right signal, which may be due to electronic crosstalk origin of the device used. The purpose of information extraction and removal of redundant data is needed to convert multi-channel signal in a different way than is offered by the manufacturer of DVR. It was assumed that only one channel is interesting for analysis. For this purpose, is used removing unnecessary columns of digits representing the individual channels of audio from the matrix of the resulting analogue to digital converter. This is done in the following way:

1. Determine the average signal level in each channel;
2. If the average signal level exceeds the threshold, write a column of data;
3. If the average signal level is below the threshold, delete a column of data.

Fig. 2. Selected audio channel with desired data.

The signal level is averaged arithmetically, while the decision on the selection of the output column is taken in accordance with the formula:

$$\left[ \frac{1}{N} \sum_{j=1}^{N} x_{n+1} \right]_{1 \leq k \leq K} \leq X_{\text{cal}}$$  \hspace{1cm} (1)$$

where $$K$$ is the number of columns in input matrix, $$N$$ is the number of values in a single column, the number of samples per channel, and $$X_{\text{cal}}$$ is a threshold value calculated from the formula:

$$X_{\text{cal}} = WP \frac{1}{K} \sum_{k=1}^{K} \left[ \frac{1}{N} \sum_{j=1}^{N} x_{n+1} \right]_{k}$$  \hspace{1cm} (2)$$

where $$WP$$ is an arbitrarily chosen partition coefficient, or from the relation determining the maximum value:
\[ X_{zal} = \max \left[ \frac{1}{N} \sum_{j=1}^{N} x_{qj} \right] \] (3)

A data selection is correct, if the threshold is properly defined using the partition coefficient. This can be done manually, in which case it is likely that the output array will contain more than one audio channel, for example, for the IWP = 1 and at least two paths containing audio data and one containing a low-noise power. In the case of recording with only one microphone for further analyses suggested saving, from a pool of columns after the selection, only the one with the highest average signal level.

4. Normalization of the length of the audio recording

The resulting single-channel audio signal, after applying the presented solutions is shown in Figure 2. Minor problems at the end of the recording are the result of expiration recorded person who was registered despite the observation fixed distance between the microphone and the mouth. This is a direct effect of the directional microphone and evidence for its proper positioning of the speaker, while maintaining a relatively low degree of sensitivity, adversely affecting the level of noise in the recording. Increasing the sensitivity of recording equipment to increase the noise occurring when a recorded person is saying nothing. The maximum value of the level of voice recorder is limited because of the risk of unacceptably high current flow through the circuit device. Reducing the degree of sensitivity necessitates approximation of the microphone to mouth the recorded person, due to low signal to noise ratio. For both cases, the ratio of averaged signal to noise is smaller than for the optimally chosen in the studies by the level of sensitivity of the recorder. This level was selected by comparing the signal to noise ratio for recording one person with three settings of the recorder in the same conditions, ambient sound and the same distance from the microphone to mouth of the recorded person. This distance is 20 +/- 5 cm and was used for all recordings.

It was observed that these data contain a large number of values close to zero at the beginning and end of recordings, which may introduce errors into the final conclusions and cause prolongation of the analysis. In addition, it should be taken into account the influence of emotions that distort the results. One of the safeguards used during the test is to prepare two voice recordings in one session. The first is rejected as potentially containing many distortions resulting from the stress during the recording process. During the registration process a person becomes accustomed to the recording equipment and the investigator and thus is prepared for the next recording. The second recording is subjected to the procedure to remove noise and silence, along with some appropriate signal at the beginning and end. This procedure is done in the following way:

1. Divide the recording into segments and sets the average amplitude in each segment;
2. If the average signal level exceeds the threshold, include a segment to the resulting matrix;
3. If this is the first segment of qualifying, remove this segment;
4. If the average signal level is below the threshold, delete the segment.

The length of the threshold should be dependent on the number of entries in the column, in matrix notation of the signal, in one second, and thus the sampling frequency. Analysing the data blocks of variable length concludes that satisfactory results can be achieved when operating at 330 ms long segments. Thus defined block length is not sufficient to recordings in which a person says phone “a” in maximum of 500 ms with intervals to 400 ms. For such cases has been successfully applied to the block length of 67 ms. The segment is then calculated using the following notation:

\[ L_{\text{seg}} = \frac{f_p}{f_s} T_{\text{def}} \] (4)

where \( T_{\text{def}} \) is the desired length of the segment in seconds and \( f_p \) is sampling frequency. In the following we describe the results for analyses using the shorter of the set of blocks. For frequencies 96 000 Hz is the value of 6400 samples in the segment. The manner of choosing is dependent on data values, which will be compared with the average of each segment. Average is appointed of the entire course, then it is divided by 6. The result is sought, determined according to the signal, a threshold value. The effect of these steps, achieved vector worthless of zero, noise and low power, which is the appropriate data set for analysis of signal and is still a sound recording. Waveform as separate data is presented in Figure 3.

[Fig. 3. The signal after removing the low-amplitude noise.]

In the last stage of the recordings chosen for further analysis is its middle section with a length of 667 ms. This process is intended to eliminate errors caused by instability of the human voice during a few
seconds of speech and regulate the length of the analysed data for all recorded people. Portions of the recordings are selected from the middle course of time, because in this point the voice is stable and is not burdened with emotional or inertia of the executive apparatus. This instability is due to changes in air pressure in the lungs, usually a decrease in pressure, which is heard as the voice of reducing the frequency and volume during prolonged speaking. The signal presented in Figure 4 is an example of selected recording. It was assumed that the person did not modulate the voice recording in a purposeful.

Fig. 4. The signal after cutting and normalization of length.

5. Normalisation of the maximum amplitude of the waveform

It was found that when analysing the fragments of recordings in which the signal level has reached the maximum level of signal range to the recording equipment, regardless of the sign on the course of time, you can not get a clear result of the frequency characteristics. This is due to the compensation due to the recorded signal of transferred voltage cut-off level above the threshold, resulting in drastically reducing the difference in level between the different resonant frequencies. In turn for recording containing quiet data obtained from all analysis were distorted. After many analyses of a variety of recordings has been proposed that suitable for further analysis are only parts of the recording of fixed length, selected from recordings with a signal level from 80% to 90% of the maximum level of range recording equipment and containing up to 2% of the fragments of silence. Suggested normalisation of the amplitude waveform of the speech signal to a value of 80% of the maximum level of range recording device via recursive multiplication of input speech and the sharing of the speech signal values exceeding 90% of the response. These operations are performed for the target data block, and the limit is the maximum value in the block division. If the signal amplitude reached the maximum level of electrical signal range for the recording equipment those parts of the recording were rejected. This allowed to minimize the error arising from the difference in amplitudes recorded voice for all people. Resulting signal, ready for the next analysis, is presented in Figure 5.

Fig. 5. Signal after normalisation of the amplitude, absolute value of the signal is between 0.8 and 0.9.

The changes observed in the results of signal processing relates solely to the shape of characteristics of the voiced vowel “a”, but not the maximum value of the signal. Doing so also allows you to compare the results with each signal analysis without the need for further normalization on them.

6. Summary

The developed procedure allows the extraction of relevant data to obtain numerical values minimally burdened by errors related to the process of recording, and emotions. In addition, reducing the number of values in the recording will accelerate further analysis. It should be noted that in the case of a single recording of many speech signals separated by silence or noise, the selected data will be free from undesirable elements. In addition, it eliminates, at the stage of data preparation, recording too short or too quiet in relation to ambient noise. It follows that the time shown on the final characteristics of the data prepared for analysis is to determine the approximate how much of the recording is interesting in comparison with the length of the whole record. In the example, from the 2.5 s recordings obtained 0.73 s, which means that 1.77 s were unnecessary, and undesirable from the standpoint of the researcher. Since then, the data length subjected to analyses is determined in a number of individual values, in which the distance between the values of the signal are consecutive 10.47 μs in the time scale on the abscissa. It should be emphasised that successive samples do not have to be so located in the original sound file (the effect of cutting elements deemed necessary), a fact that still does not affect the final results. For a block length of 67 ms may receive interference from fragmentation of data several times greater than is the case of a block of 330 ms. After analysing the graphical results of signal analysis it was found that interference will not affect the results of the speech signal.
Bibliography


Authors:

PhD Damian Krzesimowski
Kielce University of Technology
Al. 1000-lecia PP 7
25-314 Kielce
email: damiank@tu.kielce.pl