We are IntechOpen, the world's leading publisher of Open Access books
Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Effective Pitch Value Detection in Noisy Intelligent Environments for Efficient Natural Language Processing

Damjan Vlaj, Andrej Žgank and Marko Kos

Abstract

The performance of applications based on natural language processing depends primarily on the environment in which these applications are applied. Intelligent environments will be one of the major applications used to process natural language. The methods for speaker's gender classification can adapt and improve the performance of natural language processing applications. That is why, this chapter will present an effective speaker's pitch value detection in noisy environments, which then allows more robust speaker's gender classification. The chapter presents the algorithm for the speaker's pitch value detection and performs the comparison in various noisy environments. The experiments are carried out on the part of the publically available Aurora 2 speech database. The results showed that the automatically determined pitch values deviate, on average, only by 8.39 Hz from the reference pitch value. A well-defined pitch value allows a functional speaker's gender classification. In this chapter, presented speaker's gender classification works well, even at low signal to noise ratios. The experiments show that the speaker's gender classification pitch performance at SNR 0 dB is higher than 91% when the automatically determined pitch value is used. Speaker's gender classification can then be used further in the processes of natural language processing.

Keywords: intelligent environment, pitch, speech processing, gender classification

1. Introduction

Human-computer interfaces (HCIs) are frequently those parts of modern information and communications technology (ICT) systems, which play a crucial role in the case when the products are entering the market [1]. HCI is, from a user's perspective, perceived as the entity which can control the system's functionality and, thus, improve the quality of experience [2]. One of the ICT systems where HCI has made significant development progress in the last decade is the smart home and smart city solution [3, 4]. Human communication with these systems can be carried out in the form of the spoken interaction, which is the most natural and frequent modality for users. Some of the commercially available spoken virtual agents are Alexa by Amazon, Siri by Apple, Google Now by Alphabet, and Cortana by Microsoft. These commercial virtual agents support major languages but are lacking

support for under-resourced languages [5]. Another shortcoming is the lack of support for real spontaneous and emotionally driven conversation, which could improve the quality of the experience further [2].

Spoken virtual agents are using various natural language processing (NLP) techniques to communicate efficiently with humans (**Figure 1**). In this case, the NLP algorithms depend on the text input, which is produced via automatic speech recognition (ASR). Besides the direct text input, the ASR and its submodules can provide the NLP with additional meta-information, which can be used to improve the virtual agents' response to the user's communication. Some categories of such meta-information are emotions, stress level [6], effects of spontaneous speech, speakers' change [7]. As an example, change of speaker influences dialogue modelling, which can be seen as an essential part of language generation with NLP approaches. Advanced ASR systems can apply spontaneous speech modelling to reduce the ratio of errors produced by a subsequent NLP system, which has to process such an error-prone spontaneous input. Another example is the case when the NLP system is a part of an eHealth solution, where changes in stress level alter the NLP response of the virtual agent directly, either in the way of triggering some relaxing scenarios or forwarding this data in the form of NLP-generated information to caregivers [3].

An important characteristic of the human speech signal is pitch value, which can be used for both of the cases of ASR-NLP interaction mentioned above [8]. Pitch value can be used as one of the parameters, obtained during the feature extraction, which is the first step of speech recognition. Particularly, tonal languages (i.e. Mandarin) are such where pitch value information plays a crucial role in an ASR system. Moreover, the pitch value can improve speech recognition accuracy significantly in the case of spontaneous and accented speech, which is common in real-life human interaction with virtual agents. Pitch value can also be included as part of meta-information for NLP approaches, as it can be used to estimate emotions,
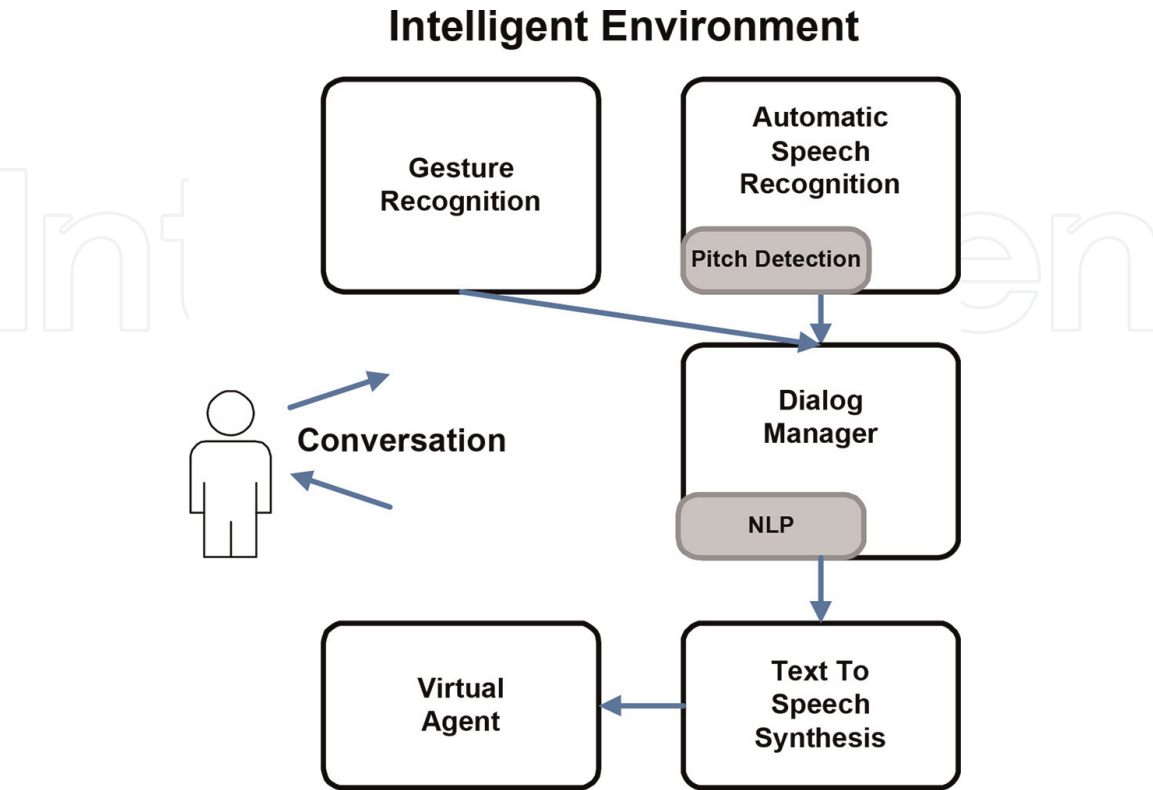

Figure 1.
Diagram of conversational human-computer interaction in an intelligent environment.

stress, and speaker information. The objective of this chapter is to present the importance of pitch value calculation from the point of view of automatic speech recognition (ASR) as a first building block of a natural language processing-based spoken virtual agent. Improved pitch value calculation can have a significant impact on the performance of an NLP system and, thus, also, in general, improve the quality of experience of virtual agents or other systems based on spoken human-computer interaction.

This chapter is organised as follows. The literature review is given in Section 2. This chapter will give a general view of the determination of the pitch value in the speech in Section 3. The pitch value determination method is presented in Section 4. The experimental design and results are given in Section 5. This chapter concludes with the discussion in Section 6 and conclusion in Section 7.

2. Literature review

Effective pitch value estimation (or fundamental frequency F_0, as it is also referred to) for various tasks and applications has been addressed for many years. It is one of the fundamental problems in speech processing because pitch value estimation is used in several different applications (e.g. speech recognition, speech perception, speech transformation, language acquisition, speech analysis, speaker identification) [9]. Good review work regarding pitch value extraction was presented by Gerhard [10]. Pitch value can be estimated in the time, spectral, or cepstral domains. It can also be extracted using auditory models. One of the time domain approaches was presented in [11]. For pitch detection, the authors proposed a so-called Yin estimator. The inspiration was the yin-yang philosophical principle of balance, which represents the authors attempt to achieve a balance between autocorrelation and cancellation, which are both implemented in the proposed algorithm. The problem of using the autocorrelation approach for pitch value estimation is that the peaks can also occur at sub-harmonics. Because of this phenomenon, it is difficult to determine which peak represents the actual fundamental frequency and which is a sub-harmonic. Yin estimation tries to solve these problems by using a difference function, which attempts to minimise the difference between the original waveform and its delayed copy. Another time domain approach is presented in [12]. The proposed approach is dealing with time domain pitch value estimation for telephone speech. Telephone speech is specific because it has reduced bandwidth and, consequently, the fundamental frequency can be very weak or even missing. In such circumstances, traditional methods based on autocorrelation cannot provide good results. To reduce the effect of narrower bandwidth, the authors propose a nonlinear filter which restores the weak or missing frequency band. After that, the combined autocorrelation function is calculated based on the original and nonlinearly processed speech. Results show 1% improvement for clean studio speech and 3% improvement for telephone speech. The experiments were performed on a Keele pitch database [13].

Pitch value can also be derived in the spectral domain, where one of the popular principles for this task is the use of tuneable filters. In [14], the author presented a method based on a narrow user-tuneable band-pass filter, which is swept across the frequency spectrum. The fundamental frequency is detected when maximum value is present on the output of the filter. The F_0 is then equal to the central frequency of the filter. The author of the paper also suggests that the difference could be detected between an evenly spaced spectrum and a richly harmonic single note. Another method using multiple comb-filter approaches was presented in [15]. The authors are investigating the problem of multiple fundamental frequencies estimation in a

noisy environment. This can happen when many persons speak at the same time with the presence of background noise. Their work is done for two speakers. The pitch value of the first speaker is determined by detecting the autocorrelation of the multi-scale product (AMP) of the mixture signal. After that, a multiple comb filter is applied to filter out the dominant signal. A residual signal is obtained after the subtraction of the remaining signal from the mixture signal. Next, the AMP is applied to the residual signal to estimate the pitch value of the second speaker. Results of the proposed method show that the method is robust and effective. Experiments were performed on the Cooke database [16]. The pitch estimation algorithm, which is robust against high levels of noise, called PEFAC, was proposed by Gonzales and Brookes [17]. The algorithm is able to identify voiced frames and estimate pitch reliably, even at negative signal-to-noise ratios. The proposed principle uses nonlinear amplitude compression to reduce narrowband noise for more robust pitch estimation. Two Gaussian mixture models (GMMs) are trained for voiced speech detection and are used for voiced/unvoiced speech classification. The proposed algorithm was evaluated on a part of the TIMIT database and on the CSLU-VOICES corpus and compared with other widely used algorithms. The tests show better performance, especially for negative SNR. The authors in [18] proposed robust harmonic features for classification-based pitch estimation. The proposed pitch estimation algorithm is composed of pitch candidate generation and target pitch selection stages. Two types of spectrum are proposed for extracting pitch candidates. One is the original noisy long-term speech spectrum, and the other is the long-term sub-harmonic summation (SBH) spectrum. If the SNR is low in the part where the F_0 is present, the F_0 spectral peak could disappear. In this case, SBH serves as a complementary source for pitch candidate extraction. In the second step of the proposed algorithm, pitch candidate classification using a neural network is performed, based on multidimensional pitch-related robust harmonic features. The five proposed features are based on the energy intensity and spectrum envelope properties of the speech. Experiments were performed on the Keele database and CSTR database. Performance of the proposed algorithm was tested against five of the common pitch estimation algorithms, including SAcC, JinWang, PEFAC, RAPT, and Yin. The results show better performance than the compared algorithms across various types and levels of noise.

Another domain where methods for pitch value extraction exist is the cepstral domain. A cepstrum is a form of spectrum where the output is the Fourier transform of the logarithmic spectral magnitude of the original waveform. The author in [19] proposed a method which needed a jury of experienced listeners for pitch value estimation judgement. Cepstrum was computed digitally and then transformed on microfilm by plotter. The method was proposed in 1967, when computer use for processing was still minimal. Another cepstrum-based method for fundamental frequency estimation was presented in [20]. Pitch information is extracted using a modified cepstrum-based method, after which the cepstrum is refined using a pitch value tracking, correction, and smoothing algorithm. In the presented work, a cepstrum-based voicing detector is also discussed. Voicing decisions are made using a multi-featured voiced/unvoiced (V/UV) classification algorithm, based on statistical analysis of the zero-crossing rate, energy of short-time segments, and cepstral peaks. Experiments were performed on speech data taken from TIMIT database. Results show considerable improvement relative to the conventional cepstrum methods. The proposed algorithm also tends to be robust against additive noise.

Pitch value can also be derived using auditory models, as the author presented in [21]. They proposed a multi-channel pitch determination algorithm (PDA), which is composed of an automatic channel selection module and a pitch value extraction module that relies on the pseudo-periodic histogram for the pitch value search.

The proposed PDA outperformed the reference system (auditory modelling AMPEX) for 0 dB SNR telephone speech and car speech. The automatic selection of channels was effective on the very noisy telephone speech but performed less successfully in the car speech. Another model-based approach was proposed by Shi et al. [22]. Their approach uses Bayesian pitch tracking, which is based on the harmonic model. Good robustness against noise was achieved by using the parametric harmonic model. A fully Bayesian approach was applied to avoid overfitting of the first-order Markov chains. Results show that the proposed algorithm has good robustness against voicing state changes, as it carries past information on pitch over the unvoiced or silent regions. Experiments were performed on Keele and Parkinson's disease databases.

Amongst other things, pitch estimation is a very important feature for the speaker's gender classification, as it is one of the more distinguishable properties between male and female speakers. Information about the speaker's gender is useful for tasks like speaker clustering or demographic data collection. In work presented in [23], the author used formant and energy-based features and several different pitch-based features for speaker's gender classification on emotionally coloured speech. Some of the features used are min., max., average pitch values, interquartile pitch range. Experiments were performed on the Danish emotional speech (DES) database, Sahad emotional speech (SES) database, and German emotional speech (GES) database. A probabilistic neural network (PNN) is a feedforward neural network, which is widely used in classification, support vector machines (SVM), K-nearest neighbour (K-NN), and GMM were compared for classification performances of naive Bayes. Results show over 90% gender classification accuracy, where the SVM classifier gave the best results.

The next section will give a general view of the determination of the pitch value in the speech signal and present what needs to be done to determine the pitch value at all.

3. Generally on the pitch value determination

Determining the speaker pitch value from the captured audio signal is possible, both in time and frequency domain representation of the signal. Determination is possible only in parts of the signal that contain a voiced speech signal. Representatives of the voiced speech signal are vowels, diphthongs, and semivowels that contain much more energy than the consonants, which are also present in the speech signal. **Figure 2** shows the time and frequency domain presentation of the vowel /eh/ and the consonant /s/ of the word "seven" in the captured audio signal. There is a considerable difference in the amplitude between the vowel /eh/ and the consonant /s/ in both the time and frequency domains. The amplitude of the vowel /eh/ (**Figure 2(a)**) is about 100 times greater than the consonant amplitude /s/ (**Figure 2(c)**). **Figure 2(a)** also has a well-seen repetitive signal pattern from which it is possible to determine a pitch value, whilst, for **Figure 2(c)**, this cannot be said. In the time domain, the pitch value or fundamental frequency F_0 can be calculated as:

$$F_0 = \frac{f_{samp}}{\tau},\tag{1}$$

where f_{samp} is the sampling frequency of the captured audio signal, and τ is the difference between the peaks. The determination of the last value is presented in **Figure 2(a)**. In this case, the value τ is (133−96) = 37 samples, whilst the sampling frequency is 8000 Hz. From this, it follows that the pitch value is equal to 216.2 Hz. The determination of the pitch value in the frequency domain is
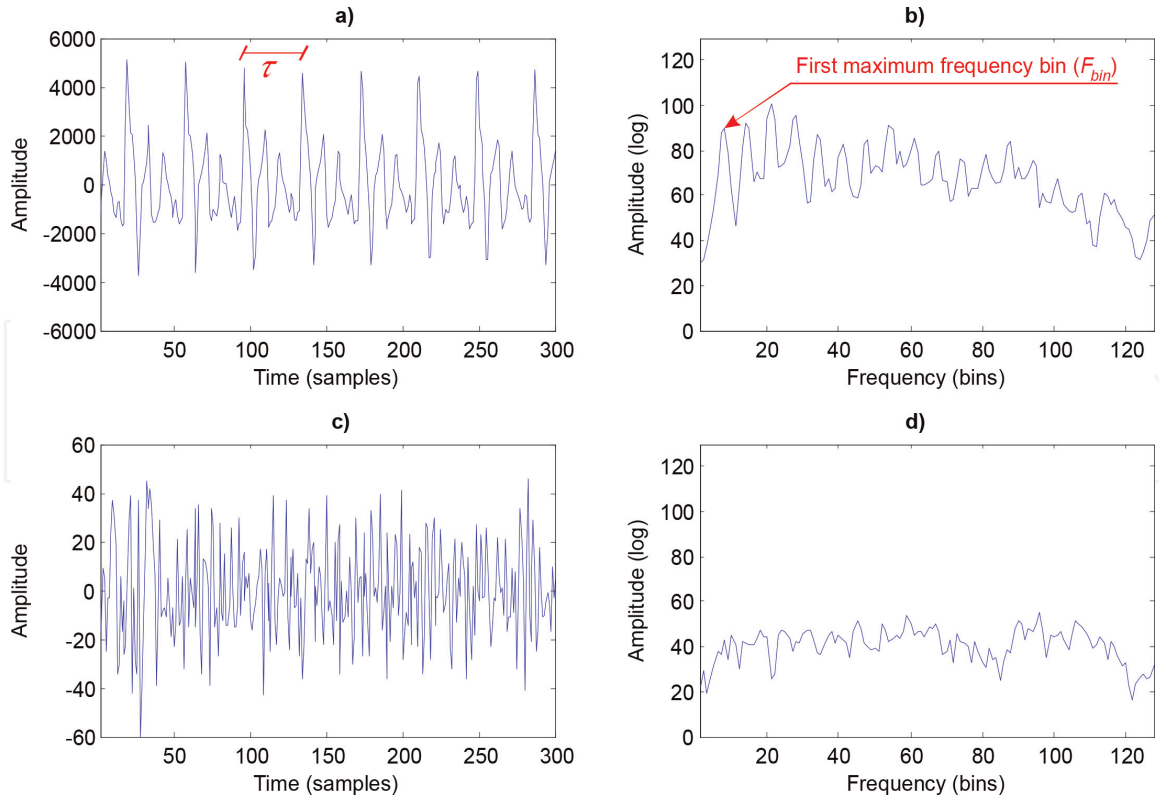
Figure 2.
Time and frequency domain presentation of the vowel /eh/ and the consonant /s/ of the word "seven" in the captured speech audio signal. (a) Voiced vowel /eh/ in time domain, (b) voiced vowel /eh/ in frequency domain, (c) unvoiced consonant /s/ in time domain and, (d) unvoiced consonant /s/ in frequency domain.

presented in **Figure 2(b)**. The pitch value can be determined by detecting the first maximum value on the frequency axis, and it is calculated as:

$$F_0 = F_{bin} \frac{f_{samp}}{2 \times F_{allbins}}, \tag{2}$$

where f_{samp} is the sampling frequency of the captured audio signal, F_{bin} is the first maximum value on the frequency axis, and $F_{allbins}$ is the number of all bins on the frequency axis. In **Figure 2(b)**, the value F_{bin} is 7 bins, the value $F_{allbins}$ is 128 bins, and the sampling frequency is 8000 Hz. So, it follows that the pitch value is 218.8 Hz. The difference between the two calculated pitch values on the same frame of the speech signal is less than 3 Hz. In the areas of the speech signal, where the consonants are located, it is not possible to determine the pitch value. Therefore, it is very important to define the boundaries of the voiced signal correctly in the whole speech signal. The voice activity detection (VAD) algorithm determines the presence of a voiced speech signal.

The VAD algorithm usually detects the presence of the entire speech signal in the captured audio signal. Such a solution is used in ASR systems to improve speech recognition accuracy. In methods for pitch value extraction, however, it is important that the VAD algorithm detects parts that contain only voiced parts of the speech signal. **Figure 3** shows the result of the voiced speech detection for the word "seven", which is obtained with the VAD algorithm.

Once the voiced parts of the speech signal are defined, then the determination of the pitch value can be made on these parts of the signal. However, another problem occurs when detecting the pitch value of a particular speaker. The pitch value of the speaker changes through pronunciation and is not constant at all times. If, as an
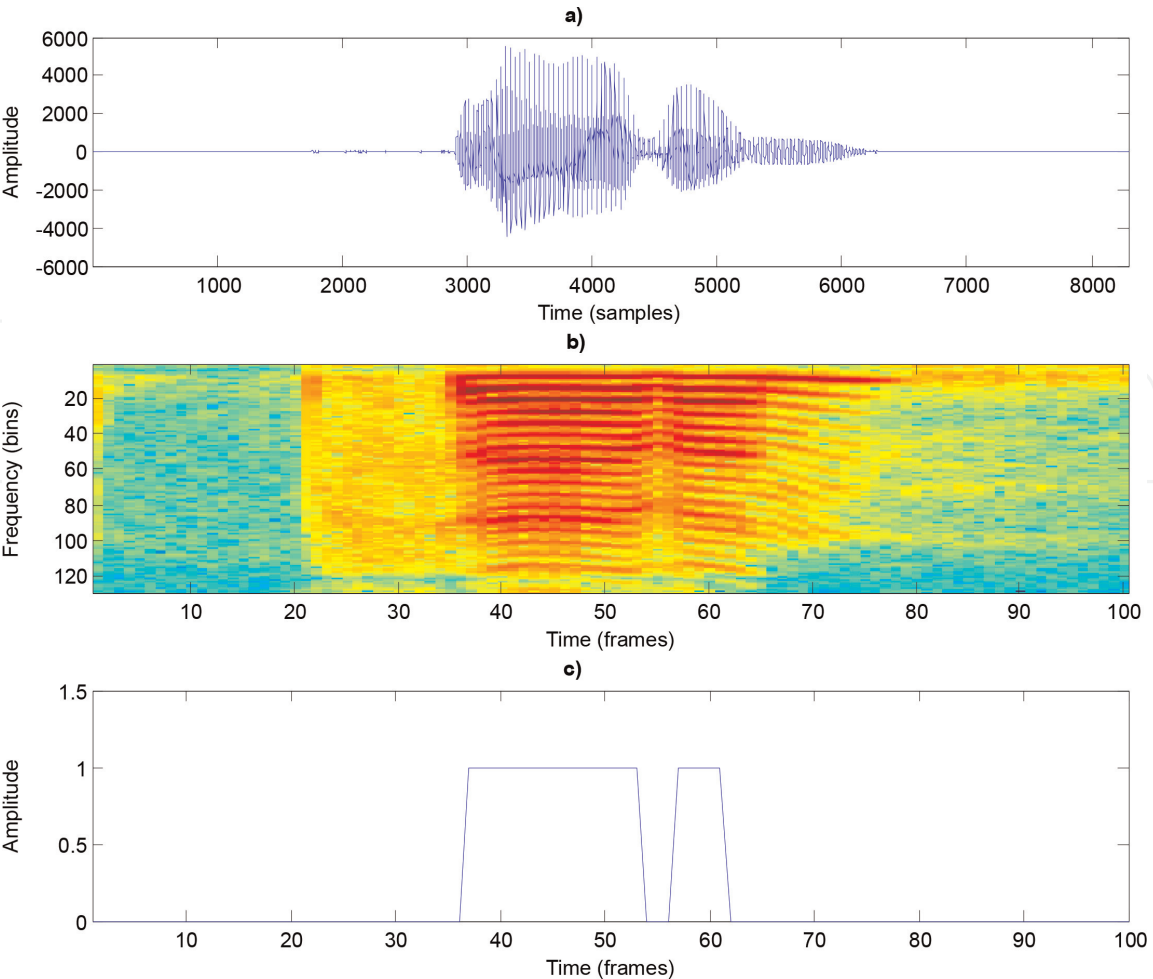
Figure 3.
VAD detection for only voiced parts of the speech signal for the word "seven": (a) audio signal in the time domain, (b) audio signal in the frequency domain, and (c) VAD decision.

example, the pronunciation of the word "seven" is taken, which is presented in **Figure 3**, the phonetic record of the word "seven" is /s eh v ah n/. In the word "seven", there are two vowels, namely /eh/ and /ah/, which represent the voiced part of the speech, in which the pitch value can be determined. For vowel /eh/, the pitch value of 216.2 Hz is determined, whilst for the vowel / ah /, the value is 222.2 Hz. As can be seen, the pitch values differ although they are very close. Because these vowels are very similar, there are no significant differences. Otherwise, there is a difference by the pronunciation of the word "zero". The phonetic record of the word "zero" is /z ih r ow/. The pitch values are determined on the vowels /ih/ and /ow/. The first vowel has a pitch value of 266.6 Hz, whilst the other has a value of 190.5 Hz. As can be seen, there are substantial differences between the calculated values although the vocals are part of one word spoken by one speaker. Such significant differences for isolated words occur but not often. However, even more significant fluctuations in pitch value detection occur in longer sentences, as the speakers, in particular, of a declarative sentence, start to speak loudly and more quietly towards the end of the sentence. This type of speech, however, contributes to a greater fluctuation of the pitch value for one speaker. Therefore, in the presented chapter, the comparative tests are made on short words, such as isolated digits.

The next section will present the process of determining the pitch value, which works well in different noise environments and also for low signal-to-noise ratios (SNRs).

4. The proposed pitch value detection procedure

As already mentioned, when determining the pitch values, it is primarily necessary to determine where the voiced speech areas are in the captured audio signal. The next subsection will present the VAD algorithm that was used to detect the voiced part of the speech for the pitch value determination process. After that, the description of the procedure will be presented which defines the pitch value in the voiced areas of the speech signal.

4.1 Voice activity detection algorithm

The voice activity detection (VAD) algorithm could have an improper effect on the results of pitch value detection. For this reason, the boundaries are determined of the beginning and end of the presence of a voiced signal in a speech signal on a clean signal only, without the presence of noise. The resulting boundaries were then also used for audio recordings with added different noisy signals with different SNR values. In order to explain the process of determining the VAD decision on the clean speech signal, **Figure 4** will be used, in which the audio signal is presented in the time domain (**Figure 4(a)**) and the frequency domain (**Figure 4(b)**) for the spoken word "four". **Figure 4(c–e)** present frame energy values and zero-crossing measure values with the corresponding threshold values.

The values of frame energy E_f are presented in **Figure 4(c)** as a blue line. The value of frame energy E_f is calculated as:

$$E_f = \frac{\sum_{i=1}^{N} s[i]^2}{N} \tag{3}$$

where $s[i]$ is a sample of an audio signal, and N the number of samples in the frame, which, in our case, is 300 samples. The energy threshold E_{th} (the red line in **Figure 4(c)**) is defined for a whole audio signal and is calculated as:

$$E_{th} = \frac{\max\left(E_f\right) + \min\left(E_f\right)}{2} \tag{4}$$

where $\max\left(E_f\right)$ is the maximum frame energy value, and $\min\left(E_f\right)$ is the minimum frame energy value in the whole audio signal.

The frame zero-crossing measure value, denoted with ZC_f, presents how many times the signal in the frame crosses the value zero or changes the sign. The zero-crossing measure value gives us additional information for the VAD decision, since it is widely known that a large zero-crossing measure value in the frame represents the frame that contains noise or frame in the audio signal which contains unvoiced speech. For example, phoneme /f/ is a consonant, which belongs to unvoiced speech. **Figure 4(d)** shows the frame zero-crossing measure value (blue line). It can be concluded from **Figure 4(d)** that the ZC_f values in the regions of unvoiced speech and noise signal are indeed large and much larger than those in the region of a voiced speech signal. The zero-crossing threshold value ZC_{th} can be set, which determines the segments of unvoiced speech and segments of voiced speech signal as:

$$ZC_{th} = \frac{\max\left(ZC_f\right) + \min\left(ZC_f\right)}{2} \tag{5}$$

where $\max\left(ZC_f\right)$ is the maximum zero-crossing measure value, and $\min\left(ZC_f\right)$ is the minimum zero-crossing measure value in the whole audio signal. The zero-crossing threshold value ZC_{th} is presented in **Figure 4(d)** as a red line.

The VAD decision is based on frame energy value E_f, frame zero-crossing measure value ZC_f, threshold energy value E_{th}, and zero-crossing threshold value ZC_{th}, as presented in Eq. (6). The proposed VAD algorithm detects voiced speech frames, since only in these frames of the speech signal, the pitch value can be determined. **Figure 4(e)** shows the VAD decision on the whole audio signal.

$$VAD_f = \begin{cases} 1; & (E_f > E_{th}) \wedge (ZC_f < ZC_{th}) \\ 0; & (E_f \leq E_{th}) \wedge (ZC_f \geq ZC_{th}) \end{cases} \qquad (6)$$
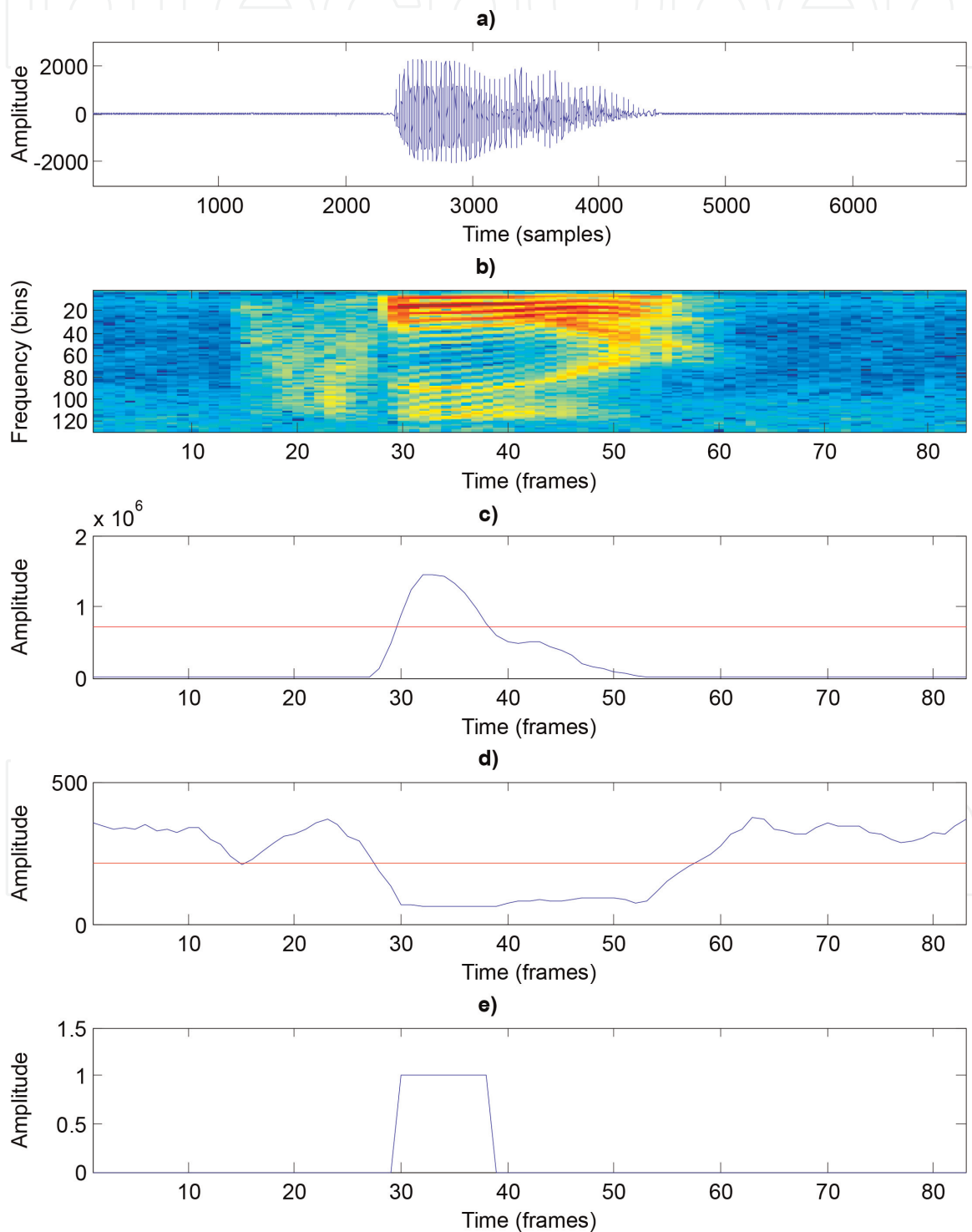

Figure 4.
VAD decision on clean speech signal for the word "four": (a) audio signal in the time domain, (b) audio signal in the frequency domain, (c) frame energy values and energy threshold value, (d) zero-crossing measure values and zero-crossing threshold value, and (e) VAD decision.

4.2 Pitch value detection

The time domain representation of the audio signal is used to determine the pitch value. The pitch value is determined in each frame which was previously detected by the VAD algorithm as a voiced part of the speech signal. To explain the process of determining the pitch value, **Figure 5** will be used, in which the part of the speech signal is presented where the word "three" is pronounced. The presentation on the vowel /iy/ will be made, which is located at the end of the pronunciation of this word.

The first step in the pitch calculation procedure is to define the highest maximum value between positive samples values and the lowest minimum value between negative samples values. The red line in **Figure 5(a)** presents the highest maximum and the lowest minimum values of the samples in the frame. The next step is to define the positive and negative peaks. Only the samples that are greater than 75% of the maximum or minimum value are used and searched for the current peak maximum or minimum. The maximums are searched in the direction of the
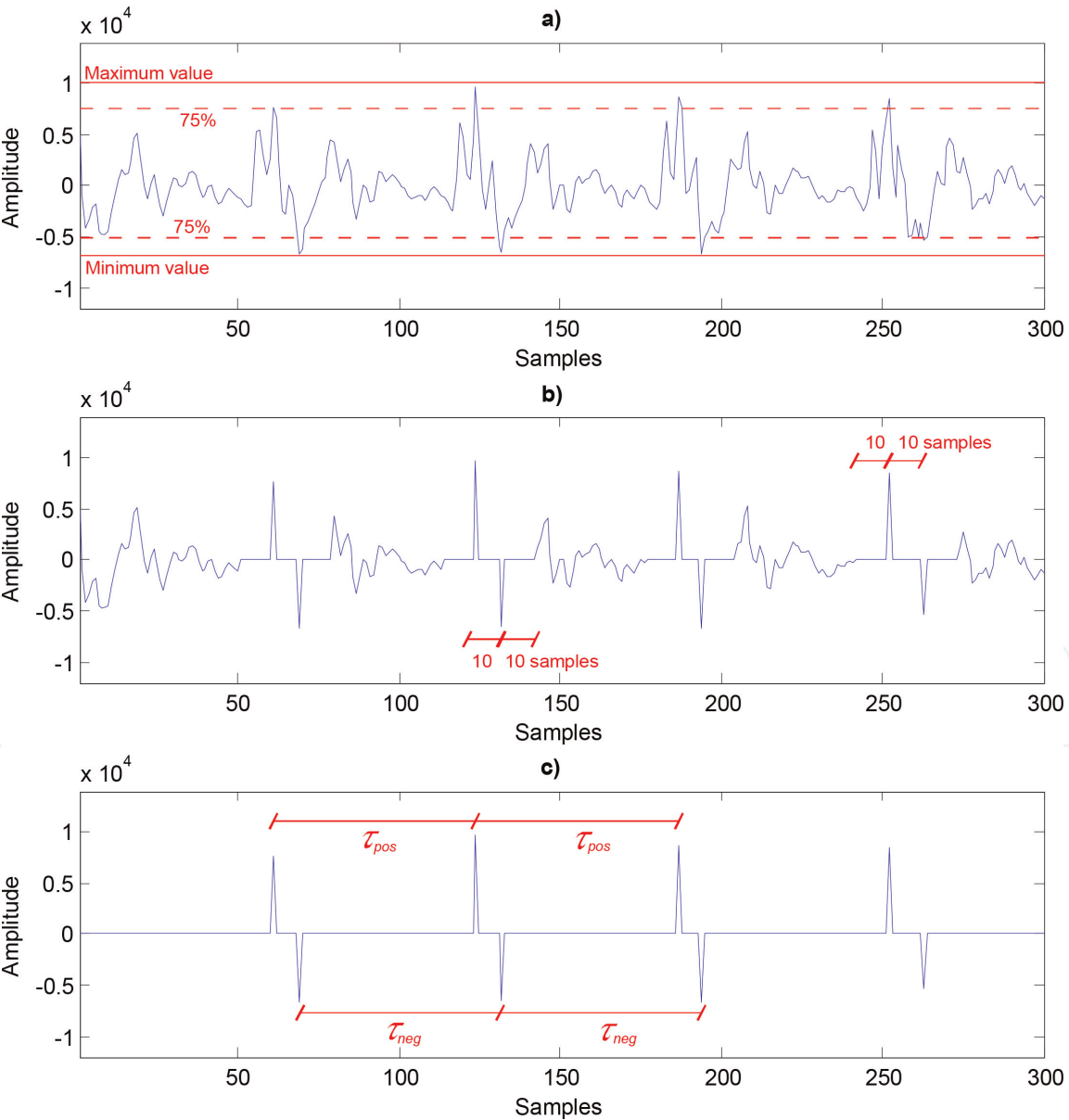

Figure 5.
The time domain representation of one frame on phoneme /ah/: (a) search for peaks in a voiced speech signal frame, (b) extraction of peaks, where 10 samples left and right around the peak are set to 0, and (c) all samples smaller than 75% of maximum and minimum values are set to 0.

highest maximum to 75% of their value. Whenever a positive or negative peak is found, the 10 samples left and right from the current positive or negative peak is set to 0. The result of this procedure is shown in **Figure 5(b)**. When all the positive and negative peaks are found, all other samples below 75% of the highest maximum or lowest minimum are set to 0. The result can be seen in **Figure 5(c)**.

The next step is to find the difference or the number of samples between the peaks. In **Figure 5(c)**, the difference is represented by the variable τ. Differences are calculated between all adjacent peaks. **Table 1** shows positive and negative peaks' positions in the presented frame and calculated differences between adjacent peaks. The difference between the last two minimum peaks is greater (look at **Figure 5(a)**); it can be seen that the minimum peak was detected incorrectly. An error occurred because the amplitude of the signal had changed slightly in that part. Just as the differences between the peaks for this represented frame can be determined, the same is done for all the frames containing the voiced speech signal. Thus, for each audio recording, a set of positive τ_{pos} and negative τ_{neg} differences is obtained between the peaks. The most commonly detected difference in each audio recording is then used to calculate the pitch value. The pitch value is determined, so that a positive pitch value F_{0pos} is obtained, with the most commonly detected positive difference between adjacent peaks and negative pitch value F_{0neg} being obtained with the most commonly detected negative difference between nearby peaks. The positive pitch value is calculated as presented in Eq. (7), and the negative pitch value as shown in Eq. (8). In Eqs. (7) and (8) the variable f_{samp} represent the sampling frequency.

$$F_{0pos} = \frac{f_{samp}}{\tau_{pos}} \tag{7}$$

$$F_{0neg} = \frac{f_{samp}}{\tau_{neg}} \tag{8}$$

If the two pitch values are the same, then the pitch value was probably determined correctly. However, if they differ, the correct pitch value is the smallest, and it is determined as:

$$F_0 = \begin{cases} F_{0pos} & F_{0pos} \leq F_{0neg} \\ F_{0neg} & F_{0pos} > F_{0neg} \end{cases} \tag{9}$$

Positive peak position	Positive difference between adjacent peaks τ_{pos}
61	
124	63
187	63
252	65
Negative peak position	Negative difference between adjacent peaks τ_{neg}
69	
132	63
194	62
263	69

Table 1.
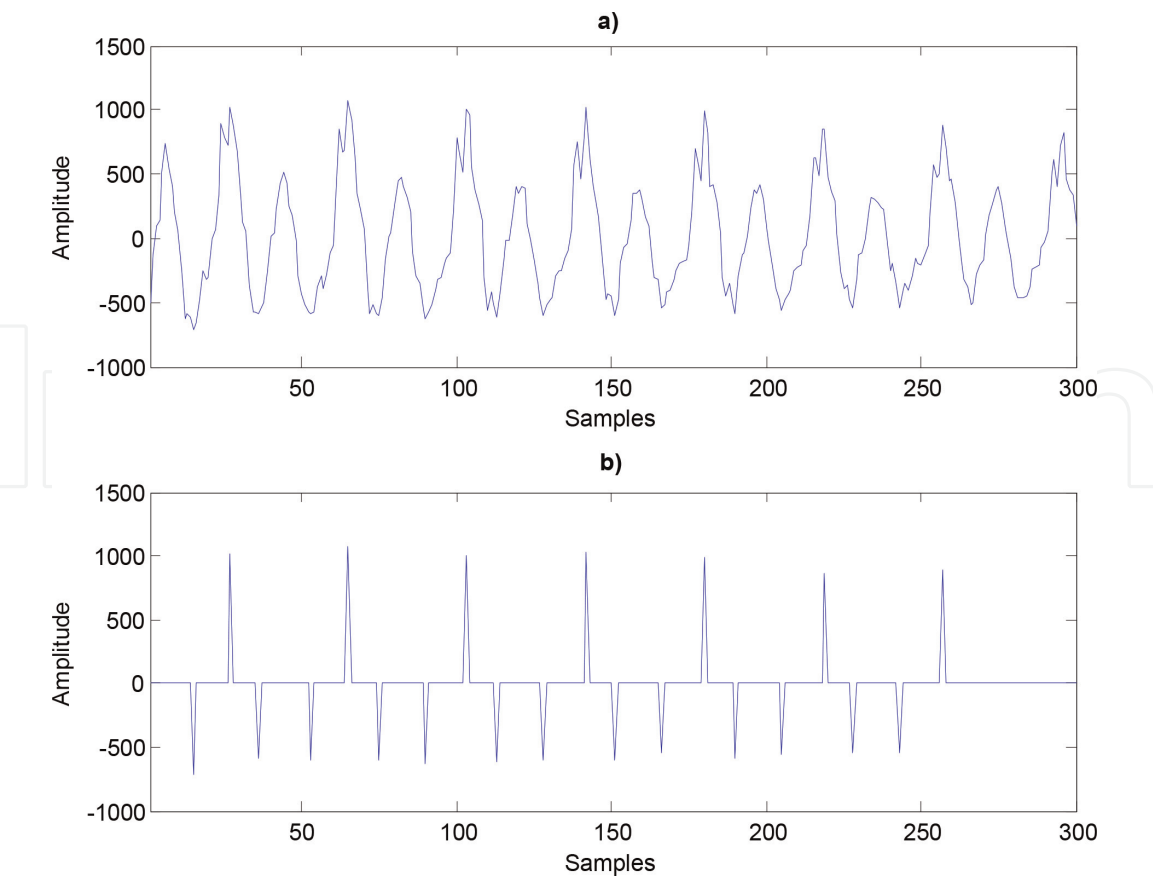*Positive and negative peaks' positions on the vowel /iy/, which is presented graphically in* **Figure 5**.

Figure 6.
Problems that may occur in peak detection: (a) time domain representation of vowel /ih/ in word "six" and (b) detected peaks' positions.

Why this decision is made can be presented in the example given in **Figure 6**. **Figure 6(b)** shows that the negative peaks are determined incorrectly. However, for positive peaks, it is evident that they are correctly defined. From this, it follows that the differences between the positive peaks are more significant than those amongst the negative peaks. To select the difference between positive peaks, consequently, the smaller pitch value is chosen.

5. Experimental design and results

In this section, the speech database is presented which was used for experiments on determining the pitch value in different noisy environments. Since the speech database used does not have reference pitch values, Subsection 5.2 will show how the reference pitch value is determined for the individual recording in the speech database. Finally, the results will be presented of the experiments on pitch value detection and gender classification.

5.1 Aurora 2 speech database

The experiments were carried out using the Aurora 2 speech database [24], which is designed to evaluate the performance of speech recognition algorithms under noisy conditions. In this chapter, the comparative tests were made only on short words, which are, in this case, isolated digits. Tests on isolated digits were chosen because, on short speech segments, the pitch value does not fluctuate so much. The speech material from the test set of the Aurora 2 speech database was

used for the experiments on pitch value detection. Three different test sets were defined for the testing. Four subsets with 298, 279, 283, and 284 utterances were obtained by splitting 1144 utterances from 52 male and 52 female speakers. The recordings of all speakers were present in each subset. Individual noise signals at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and − 5 dB were added, and the clean case without added noise was taken as the seventh condition.

The first test set is called test set A. In this test set, four noises: a suburban train, babble, a car, and an exhibition hall were added to the four subsets. The second test set is called test set B. This test was created in the same way as test set A, the only difference is that four different noises were used, which are a restaurant, a street, an airport, and a train station. The third test set is called test set C, and it contained only the first two of four subsets, with 298 and 279 utterances. Here, speech and noise are filtered using the Motorola Integrated Radio Systems (MIRS) characteristic [25], before being added to the SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and − 5 dB. The MIRS filter represents a frequency characteristic that simulates the behaviour of a telecommunication terminal, which meets the official requirements for the terminal input frequency response as specified, e.g., in the European Telecommunications Standards Institute - Special Mobile Group (ETSI-SMG) technical specification [25]. The suburban train and street were used as added noise signals. The purpose of this set was to show the influence on pitch value when a different frequency characteristic is present in the speech signal.

Both parts, training and test material of the Aurora 2 speech database, were used for the experiments on gender classification. The gender classification experiments used the same test material as the experiments on pitch value detection. As mentioned earlier, there are 1144 audio recordings in the test set. These audio recordings are divided into 570 recordings containing male and 574 recordings containing female speakers. Most of the gender classification tests were based on GMMs. The training of GMMs requires training material, which was taken from the training part of the Aurora 2 speech base. The concept of the Aurora 2 speech database experiments includes two training modes, which are defined as training on clean data only and as training on clean and noisy (multi-condition) data. From the Aurora 2 speech database, 8440 utterances were chosen for training on clean data, which contained the recordings with 4220 male and 4220 female speakers. The same 8440 recordings were used for multi-condition training. They were divided into 20 subsets, each of which included 422 utterances. The 20 subgroups represented four different noise scenarios (a suburban train, babble, a car, and an exhibition hall) at five different SNRs.

5.2 Definition of the reference pitch values

The Aurora 2 speech database does not provide information about the pitch value in each audio recording. Therefore, the reference values on 1144 audio recordings were determined manually using graphical representations of audio recordings in the time domain. The area of voiced speech in the audio recording was determined using the VAD algorithm presented in Subsection 4.1. The reference values were determined on isolated digits, which were spoken in American English. **Table 2** lists isolated digits with phonetic transcription. Phonemes written in bold represent the vowels on which the determination of the reference pitch values was made. In determining the reference pitch value, the difference between the two peaks was defined, which, as much as possible, have similar amplitude. In **Figure 4(a)**, this would be around 2800 samples. To determine the reference pitch value for a given audio recording Eq. (1) is used. The sampling frequency of audio recordings in the Aurora 2 speech database is 8 kHz.

Word (isolated digit)	Phonetic transcription
one	w **ah** n
two	t **uw**
three	th r **iy**
four	f **ao** r
five	f **ay** v
six	s **ih** k s
seven	s **eh** v ah n
eight	**ey** t
nine	n **ay** n
zero	z **ih** r ow
oh	**ow**

Table 2.
The lists of isolated digits with phonetic transcription.

5.3 Results of pitch value determination

All the presented results in this subsection were achieved with automatic pitch value determination for each audio recording. The automatic pitch value determination was based on the procedure given in Subsection 4.2. The area of the voiced speech signal was determined on a clean speech signal with the VAD algorithm presented in Subsection 4.1. The results of the VAD algorithm were used in all recordings, including those to which different noises were added at different SNR values. The present work did not use the VAD algorithm on noisy audio recordings because it is necessary to present how well the pitch value determination algorithm works, even in noisy environments. If the VAD algorithm is also used on noisy audio recordings, the VAD algorithm could have an overwhelming effect on the pitch value determination results.

Table 3 gives the results of the absolute deviation of the automatically obtained pitch value in a positive or negative direction concerning the reference pitch value.

Test set	Noise/[dB]	Clean	SNR 20	SNR 15	SNR 10	SNR 5	SNR 0	SNR −5	Average
A	Subway	1.41	2.51	2.60	3.39	4.54	9.96	24.12	6.93
	Babble	1.46	2.60	3.09	3.68	6.12	10.87	32.17	8.57
	Car	1.56	3.18	3.14	4.09	5.42	14.95	30.57	8.99
	Exhib.	1.83	2.76	3.46	3.71	4.59	10.96	24.86	7.45
B	Rest.	1.41	2.29	2.90	3.71	5.90	11.56	25.80	7.65
	Street	1.46	2.61	2.95	3.76	4.99	14.63	25.60	8.00
	Airport	1.56	2.28	2.73	3.09	6.10	11.17	24.05	7.28
	Train	1.83	2.65	3.40	3.77	7.36	11.53	25.40	7.99
C	Subway	3.49	3.64	3.57	4.86	5.56	14.96	36.78	10.41
	Street	3.73	4.10	4.27	4.53	8.07	15.52	33.88	10.58
Overall		1.97	2.86	3.21	3.86	5.86	12.61	12.61	8.39

Table 3.
Pitch value deviation results F_{odev} in [Hz] in the positive or negative direction for the individual test set according to the reference pitch value.

For each audio recording, the F_0 value was determined automatically and compared with the reference pitch value F_{0ref}. The result F_{0dev} is a deviation from the reference value presented in Hz and was calculated as:

$$F_{0dev} = \left| F_0 - F_{0ref} \right| \tag{10}$$

The test set A in the subway set has 298 audio recordings for each SNR noise level. The presented result is the average error value in Hz over all of the 298 audio recordings in each noise level. The same results were obtained for all remaining noise sets.

Tables 4–7 give the results obtained as a percentage. **Table 4** specifies how many percentages of audio recordings have the same value as the automatically obtained pitch value relative to the reference pitch value. **Table 5** shows how many percentages of audio recordings had an error between 1 and 10 Hz, corresponding

Test set	Noise/[dB]	Clean	SNR 20	SNR 15	SNR 10	SNR 5	SNR 0	SNR −5	Average
A	Subway	63.09	54.70	51.34	50.00	42.28	34.90	16.11	44.63
	Babble	72.04	60.22	56.99	50.18	41.94	32.97	16.85	47.13
	Car	63.96	56.89	53.36	49.47	43.11	28.27	16.96	44.57
	Exhib.	63.73	54.23	45.07	45.42	35.21	29.58	17.61	41.55
B	Rest.	63.09	55.37	53.69	51.01	44.63	31.88	20.13	45.69
	Street	72.04	56.63	56.99	45.88	41.22	27.96	16.49	45.32
	Airport	63.96	59.01	52.65	54.06	41.34	40.64	24.03	47.96
	Train.	63.73	54.58	50.00	45.07	41.55	33.10	19.37	43.91
C	Subway	46.31	46.31	44.63	40.27	35.23	26.17	9.40	35.47
	Street	44.44	44.44	44.09	44.44	32.62	26.52	13.98	35.79
Overall		61.68	54.30	50.94	47.66	40.00	31.26	17.10	43.27

Table 4.
Percentage of the pitch values for the individual test set that matched the reference pitch value fully.

Test set	Noise/[dB]	Clean	SNR 20	SNR 15	SNR 10	SNR 5	SNR 0	SNR −5	Average
A	Subway	36.24	39.60	42.62	39.93	45.64	43.96	41.28	41.32
	Babble	27.24	36.56	35.13	41.94	43.73	44.44	31.54	37.22
	Car	35.34	36.75	39.93	40.64	42.40	43.11	33.22	38.77
	Exhib.	35.56	41.55	48.94	45.42	53.17	48.24	37.68	44.37
B	Rest.	36.24	40.60	39.60	40.60	42.28	44.97	32.55	39.55
	Street	27.24	36.92	36.92	44.44	48.39	44.09	42.29	40.04
	Airport	35.34	37.10	41.70	37.10	42.76	38.87	28.27	37.30
	Train.	35.56	42.25	42.96	46.13	43.31	42.96	38.03	41.60
C	Subway	45.64	43.96	46.98	47.99	49.66	43.96	30.87	44.15
	Street	44.09	43.37	46.24	42.65	46.59	40.50	27.60	41.58
Overall		35.85	39.87	42.10	42.68	45.79	43.51	34.33	40.59

Table 5.
Percentage of the pitch values for the individual test set that the error of the pitch value was between 1 and 10 Hz according to the reference pitch value.

Test set	Noise/[dB]	Clean	SNR 20	SNR 15	SNR 10	SNR 5	SNR 0	SNR −5	Average
A	Subway	0.67	4.36	5.03	7.38	8.39	8.39	12.75	6.71
	Babble	0.00	1.43	4.66	4.66	7.89	11.11	11.11	5.84
	Car	0.00	3.89	4.59	5.65	7.77	9.89	12.01	6.26
	Exhib.	0.00	2.82	3.17	7.04	8.10	10.92	15.14	6.74
B	Rest.	0.67	3.02	5.37	5.37	8.05	10.74	14.77	6.86
	Street	0.00	4.30	3.58	6.45	6.45	9.68	11.83	6.04
	Airport	0.00	1.41	3.53	6.71	9.19	7.07	16.25	6.31
	Train.	0.00	1.76	4.93	5.99	7.04	10.21	8.80	5.53
C	Subway	5.03	6.71	5.70	8.05	10.40	10.74	15.10	8.82
	Street	8.60	8.60	5.73	8.60	10.75	12.90	13.62	9.83
Overall		1.50	3.85	4.65	6.61	8.43	10.17	13.18	6.91

Table 6.
Percentage of the pitch values for the individual test set that the error of the pitch value was between 11 and 20 Hz according to the reference pitch value.

to the reference pitch value. **Tables 6** and **7** give results similar to **Table 5**, but **Table 6** represents the percentage of audio recordings with errors between 11 and 20 Hz, and **Table 7** represents the percentage of audio recordings with errors above 21 Hz.

5.4 Results on gender classification

The gender classification experiments presented in this chapter will show how important the correct pitch value detection is for the gender classification. The tests were performed on the Aurora 2 voice database, and as presented in Subsection 5.1, 1144 audio recordings were available, of which 570 were with a male speaker and 574 with a female speaker. Seven experiments were performed. The first six tests

Test set	Noise/[dB]	Clean	SNR 20	SNR 15	SNR 10	SNR 5	SNR 0	SNR −5	Average
A	Subway	0.00	1.34	1.01	2.68	3.69	12.75	29.87	7.33
	Babble	0.72	1.79	3.23	3.23	6.45	11.47	40.50	9.63
	Car	0.71	2.47	2.12	4.24	6.71	18.73	37.81	10.40
	Exhib.	0.70	1.41	2.82	2.11	3.52	11.27	29.58	7.34
B	Rest.	0.00	1.01	1.34	3.02	5.03	12.42	32.55	7.91
	Street	0.72	2.15	2.51	3.23	3.94	18.28	29.39	8.60
	Airport	0.71	2.47	2.12	2.12	6.71	13.43	31.45	8.43
	Train.	0.70	1.41	2.11	2.82	8.10	13.73	33.80	8.95
C	Subway	3.02	3.02	2.68	3.69	4.70	19.13	44.63	11.55
	Street	2.87	3.58	3.94	4.30	10.04	20.07	44.80	12.80
Overall		1.01	2.06	2.38	3.15	5.87	15.14	35.49	9.30

Table 7.
Percentage of the pitch values for the individual test set that the error of the pitch value was greater than 21 Hz according to the reference pitch value.

used GMMs, and in the seventh test, the gender classification was determined based on the pitch value. Two separate models were trained for gender classification, one for the male speaker and one for the female speaker. GMMs were trained using the procedures described in [26]. The training of the GMMs was done on the training material of the Aurora 2 speech base, which contains two training modes (clean and multi-condition). The clean training material was used for the first three of the six GMMs training processes. The results of these experiments are presented in **Figure** 7. For the other three, the multi-condition content was used, the results of which are shown in **Figure 8**. The results in both figures are given as accuracy (*Acc*) in the percentage of the correct speaker gender classification, and it was calculated as:

$$Acc = \frac{H}{N} \cdot 100[\%] \tag{11}$$

where *H* is the sum of all correct classifications for male and female, divided by the number of all classifications *N*, which is 2865 for all noisy conditions.

The MFCC_E_D_A features were used for the first and fourth tests, where the entire audio recording was used for training without using segmentation. The MFCC_E_D_A features consisted of 12 Mel frequency cepstral coefficients C1-C12, logarithmic energy, and the first and second derivatives of those coefficients. The determination of the MFCC_E features is described in [27]. The procedure of calculating the first and second derivatives is described in [26]. MFCC_E_D_A features were also used for the second and fifth tests, but, in this case, segmentation was used, based on the VAD algorithm presented in Subsection 4.1. In this case, only parts of the audio recordings that contain only parts of the voiced speech signal from the audio recordings were used for training. For the third and sixth tests, an additional feature was used, namely the pitch value determined in each frame. In this case, the 12th coefficient C12 was replaced by the pitch value. So, for these two tests, the MFCC_Pitch_E_D_A features were used to train the GMM models. In this case, segmentation was also used, so that only parts of audio recordings that contained only voiced speech signals were used for training. The last, seventh test, however, was performed based on determining the pitch value F_0 for each audio recording. The results of this test are given as the last set of columns in both **Figures** 7 and **8**. The pitch value limit was set at 155 Hz, so that the speaker's gender classification was defined as:
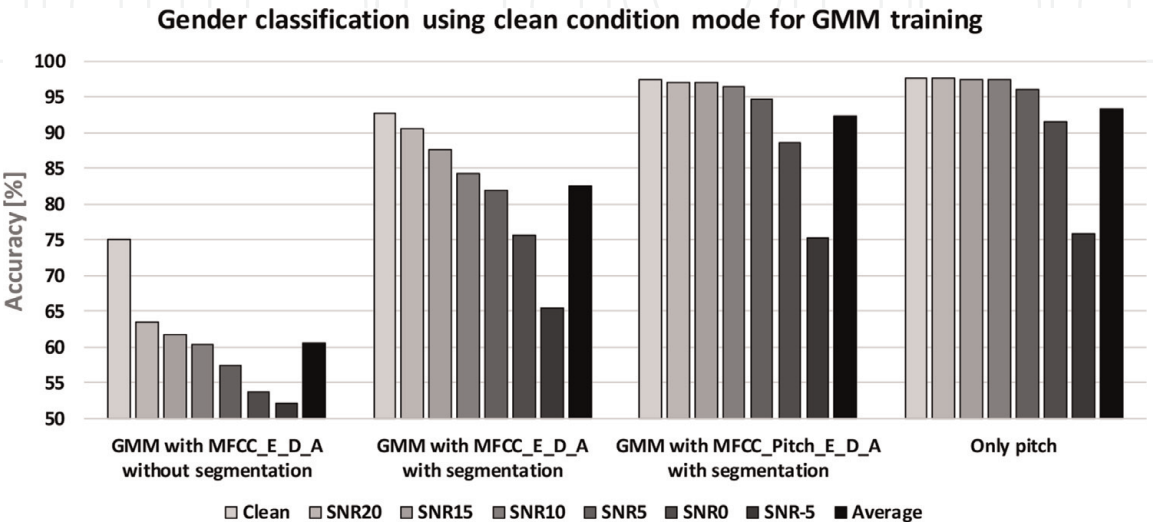

Figure 7.
Gender classification using the clean condition training mode of the Aurora 2 speech database.
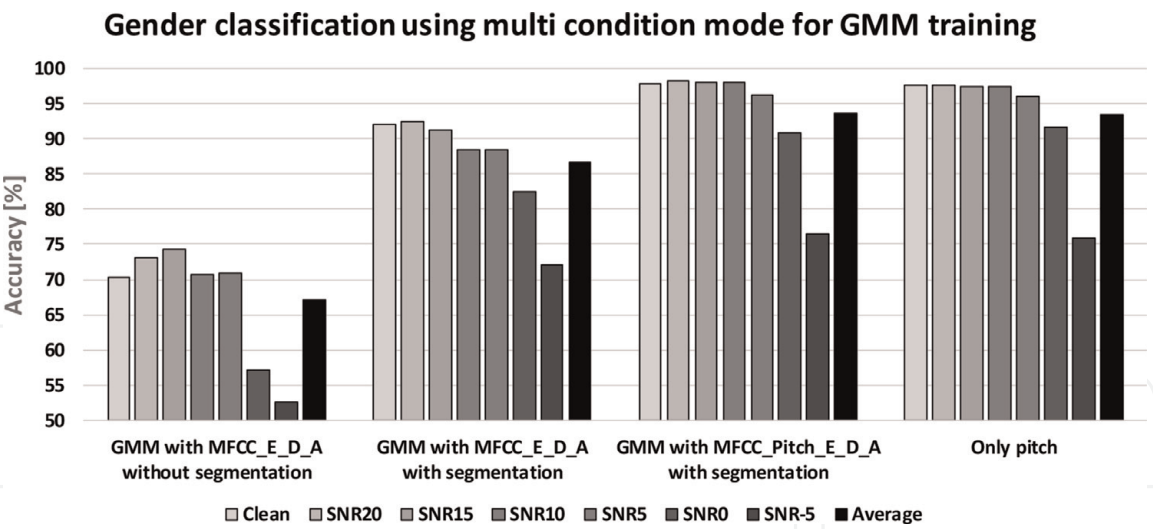
Figure 8.
Gender classification using the multi-condition training mode of the Aurora 2 speech database.

Gender	Minimum	Maximum	Average
Male	78.43 Hz	170.21 Hz	120.86 Hz
Female	131.14 Hz	275.86 Hz	206.16 Hz

Table 8.
The manual determination of the pitch values on the audio recordings obtained from the Aurora 2 speech database.

$$Gender = \begin{cases} Male & F_0 \leq 155 \\ Female & F_0 > 155 \end{cases} \tag{12}$$

The pitch value limit determination was based on the manual determination of the pitch values on the audio recordings obtained from the Aurora 2 speech database. **Table 8** gives the minimum, maximum, and average values for the pitch value for the male and female speakers, which were obtained from the manual determination of the pitch values on the audio recordings of the Aurora 2 speech database. The limits for the pitch values for a male speaker are between 78 and 171 Hz, whilst the limits for the pitch value for the female speaker are between 131 and 276 Hz. As can be seen, there is some overlapping of the pitch values. The 155 Hz pitch value limit is based on an analysis of the number of errors that could occur if the pitch for the male speaker is above 155 Hz and the pitch value for the female speaker below 155 Hz. The analysis value was also performed on the remaining pitch values between 131 Hz and 171 Hz, but the proposed pitch value limit produced the smallest number of errors.

6. Discussion

The results presented in this chapter show that the proposed automatic pitch value determination algorithm works well. For more than half of the audio recordings with the SNR higher or equal to 15 dB, the determined pitch value compared with the reference pitch value was correct (see **Table 4** for overall value). Interesting is that, on average, 17.10% full match of pitch value was achieved for audio

recordings with the SNR value −5 dB. If it is taken into account that pitch differences up to 10 Hz are still an acceptable error, then, from **Tables 4** and **5**, it can be concluded that, on average, in all noisy environments with different SNR values, the algorithm can correctly detect 83.86% of the pitch values for all audio recordings.

It is evident from **Tables 6** and 7 that, even in the case of a clean signal, there are errors greater than 11 and 21 Hz compared with reference pitch value. This is due mainly to the problem described in Section 3, since the pitch value can change during the pronunciation of some words, especially if there are several different vowels in the word. At the beginning of the word, one pitch value is determined, whilst another value can be detected at the end of the word. As described in Subsection 5.2, the reference pitch value is determined on one vowel. If the word contains multiple vowels, various pitch values can be determined. In our case, the proposed pitch determination algorithm selected the pitch value that was the most often determined from the differences between the detected peaks of the voiced speech signal.

For a clean signal, the deviation of the pitch value was, on average, below 2 Hz (see **Table 3**). As can be seen from the same table, the maximum deviation of the average values in Hz was made by the test set C. In this test set, the audio recordings were filtered with an MIRS filter, which simulates the behaviour of the telecommunications terminal. The frequency response of the MIRS filter is presented in [25]. If the values of test set A for the subway noise set and test set C for the subway noise set for a clean signal are compared, the average deviation value from the reference value is 2.08 Hz when the audio recordings were filtered with the MIRS filter. In **Figure 9**, an example of the word "five" before (test set A) and after (test set C) is filtered with the MIRS filter. As can be seen, the amplitude of the speech signal is about one-third smaller after the filter was used.

Figure 10 shows the process of determining the peaks in a voiced speech signal. As can be seen, there are errors in peak detection, especially when the signal was filtered with the MIRS filter (test set C).

Based on the good determination of the pitch value, the obtained results can be used in the gender classification. As can be seen from the results presented
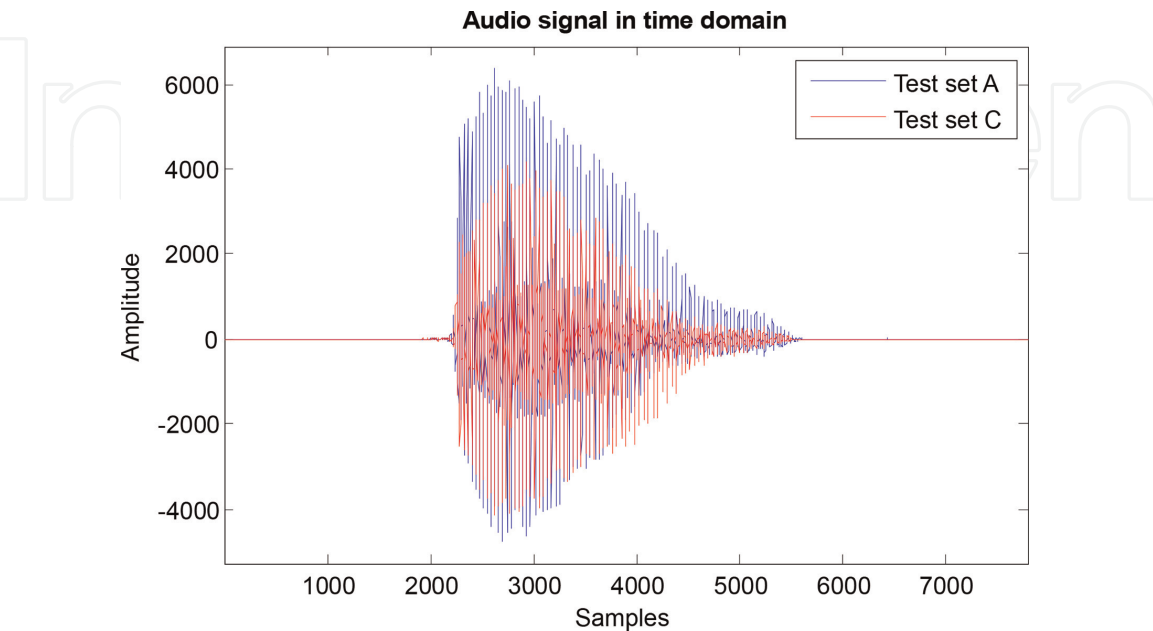

Figure 9.
The audio signal in the time domain of the word "five" before (blue line) and after (red line) it was filtered with the MIRS filter.
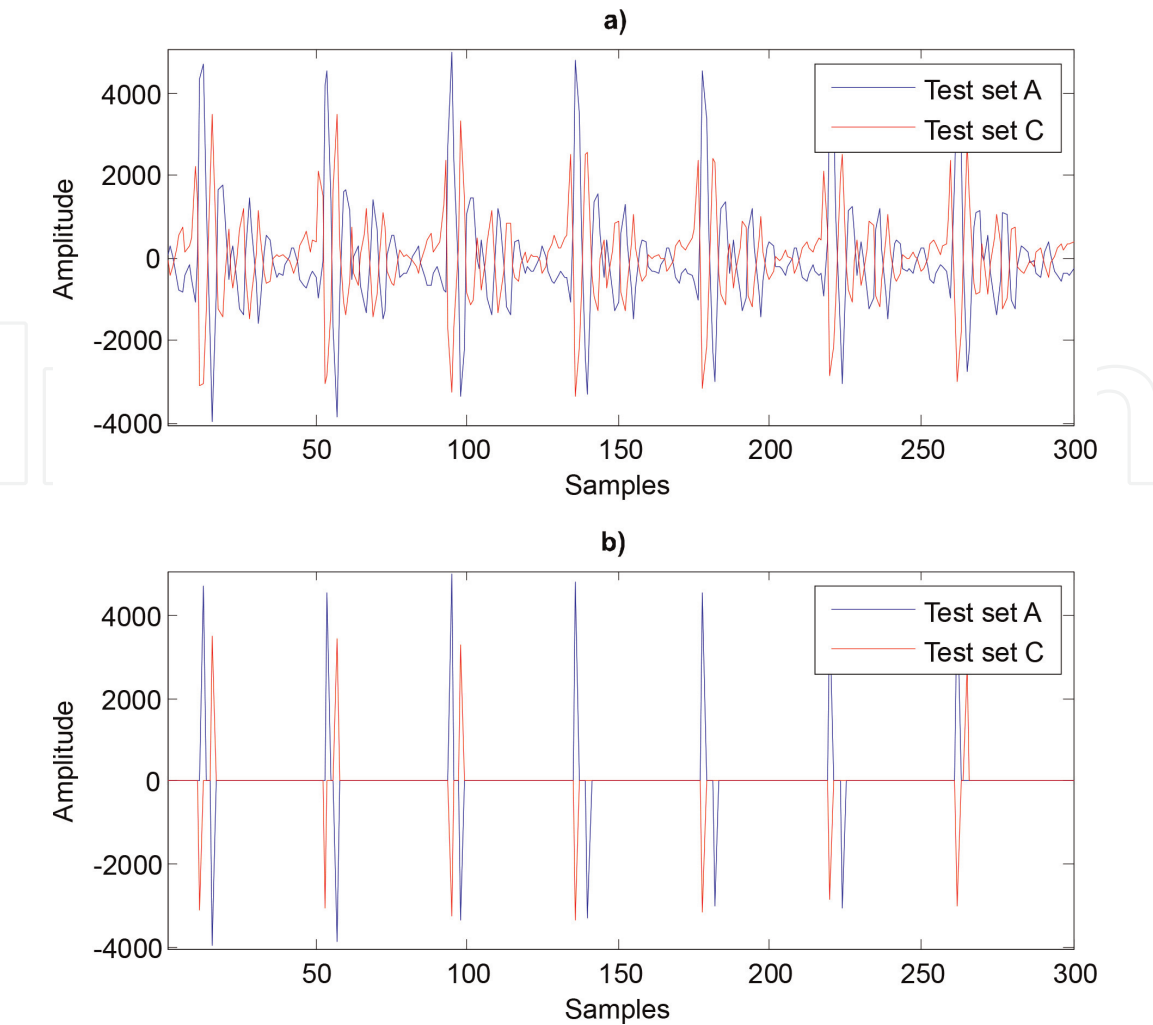
Figure 10.
The signals before (blue line) and after (red line) were filtered with the MIRS filter: (a) time domain representation of vowel /ay/ in the word "five" and (b) detected peaks positions.

in **Figures 7** and **8**, well-defined pitch values contributed significantly to the accuracy of the speaker's gender classification in both training modes (see the third set of columns in both figures). In both Figures, the fourth set of column presents results, where only the pitch value is used to classify speaker gender. The results show clearly that, even at low signal-to-noise ratios (SNR = 5 dB), the pitch value determination allowed good classification of the speaker's gender. Speaker gender classification accuracy is above 96% for SNR5. In this case, the performance was better than the performance using GMM models. However, when using GMMs, the speaker's gender classification results can be better if more training material is used. If only the pitch value is used for classification, using a different speech database will likely require a new pitch limit value to be defined. The results show, however, that the pitch value used as an additional coefficient for the features contributed greatly to the accuracy of the speaker's gender classification.

However, once a useful speaker's gender classification is made, then this can be used in intelligent environments, where the performance of natural language processing can be improved.

7. Conclusion

An effective determination of the pitch values, which works well in various noise environments, is presented in this chapter. At the beginning of this chapter,

an overview is made of the pitch values used in the technologies of natural language processing. After that, the general procedures are presented for determining the pitch value in the time and frequency domains. The main part of this chapter is the presentation of the proposed procedure for determining the pitch values. The experiments were carried out on a part of the Aurora 2 speech database. Only isolated digits were used in the tests. Isolated digits represent short words on which the pitch value can be determined without major changes during the speech pronunciation. As presented in this chapter, this may also happen in short words and even more often with longer sentences. The results showed that automatically determined pitch values for all noisy environments deviated, on average, by 8.39 Hz compared with the reference pitch values.

A well-defined pitch value allows a functional speaker's gender classification. The pitch value determination procedure presented in this chapter provides a good speaker's gender classification, even at low signal-to-noise ratios. Thus, when the automatically determined pitch value is used, the speaker's gender classification performance at SNR 0 dB is higher than 91%. A speaker's gender classification can be then used further in the processes of natural language processing.

Author details

Damjan Vlaj*, Andrej Žgank and Marko Kos
Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

*Address all correspondence to: damjan.vlaj@um.si

IntechOpen

References

[1] Pieraccini R, Lubensky D. Spoken language communication with machines: The long and winding road from research to business. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Berlin, Heidelberg: Springer; 2005. pp. 6-15

[2] Côté N, Berger J. Speech communication. In: Möller S, Raake A, editors. Quality of Experience. T-Labs Series in Telecommunication Services. Cham: Springer; 2014

[3] Vacher M, Istrate D, Portet F, Joubert T, Chevalier T, Smidtas S, et al. The sweet-home project: Audio technology in smart homes to improve well-being and reliance. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011. pp. 5291-5294

[4] Brdiczka O, Langet M, Maisonnasse J, Crowley JL. Detecting human behavior models from multimodal observation in a smart home. IEEE Transactions on Automation Science and Engineering. 2008;**6**(4):588-597

[5] Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. Speech Communication. 2014;**56**:85-100

[6] Giannakakis G, Grigoriadis D, Giannakaki K, Simantiraki O, Roniotis A, Tsiknakis M. Review on psychological stress detection using biosignals. IEEE Transactions on Affective Computing. 2019. DOI: 10.1109/TAFFC.2019.2927337

[7] Wanner L, André E, Blat J, Dasiopoulou S, Farrùs M, Fraga T, et al. Kristina: A knowledge-based virtual conversation agent. In: International Conference on Practical Applications of Agents and Multi-Agent Systems. Cham: Springer; 2017. pp. 284-295

[8] Mary L. Significance of prosody for speaker, language, emotion, and speech recognition. In: Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition. Cham: Springer; 2019. pp. 1-22

[9] Drugman T, Huybrechts G, Klimkov V, Moinet A. Traditional machine learning for pitch detection. IEEE Signal Processing Letters. 2018;**25**(11):1745-1749. DOI: 10.1109/LSP.2018.2874155

[10] Gerhard D, Pitch extraction and fundamental frequency: History and current techniques, Technical Report TR-CS 2003-06; 2003

[11] de Cheveigne A, Kawahara H. Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America. 2002;**111**(4):1917-1930

[12] Chang L, Xu J, Tang K, Cui H. A new robust pitch determination algorithm for telephone speech. In: 2012 International Symposium on Information Theory and its Applications. Honolulu, HI; 2012. pp. 789-791

[13] Plante F, Meyer G, Ainsworth WA. A pitch extraction reference database. In: EUROSPEECH'95. Madrid; 1995. pp. 837-840

[14] Lane JE. Pitch detection using a tunable IIR filter. Computer Music Journal. 1990;**14**(3):46-59

[15] Zeremdini J, Anouar M, Messaoud B, Bouzid A. Multiple comb filters and autocorrelation of the multi-scale product for multi-pitch estimation. Applied Acoustics. 2017;**120**:45-53. DOI: 10.1016/j.apacoust.2017.01.013

[16] Cooke M, Barker J. An audio-visual corpus for speech perception and automatic speech recognition. Journal of Acoustic Society of America. 2006; **120**(5):2421-2424

[17] Gonzalez S, Brookes M. PEFAC—A pitch estimation algorithm robust to high levels of noise. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014;**22**(2): 518-530

[18] Wang D, Yu C, Hansen JHL. Robust harmonic features for classification-based pitch estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017;**25**(5): 952-964

[19] Noll A. Cepstrum pitch determination. Journal of the Acoustical Society of America. 1967;**41**(2):293-309

[20] Ahmadi S, Spanias AS. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. IEEE Transactions on Speech and Audio Processing. 1999;**7**(3):333-338

[21] van Immerseel L, Martens J. Pitch and voiced/unvoiced determination with an auditory model. Journal of the Acoustical Society of America. 1992; **91**(6):3511-3526

[22] Shi L, Nielsen JK, Jensen JR, Little MA, Christensen MG. Robust Bayesian pitch tracking based on the harmonic model. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;**27**(11): 1737-1751

[23] Sedaaghi MH. Gender classification in emotional speech. In: Mihelic F, Zibert J, editors. Speech Recognition. Rijeka: IntechOpen; 2008. DOI: 10.5772/ 6385

[24] Hirsch HG, Pearce D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of ISCA Tutorial and Research Workshop (ITRW) on ASR. Paris, France; 2000

[25] ETSI-SMG Technical Specification, European digital cellular telecommunication system (phase 1)— Transmission planning aspects for the speech service in GSM PLMN system, ETSI-SMG technical specification GSM03.50, Version 3.4.0. Valbonne, France; 1994

[26] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu XA, et al. The HTK book, Version 3.4. Cambridge University Engineering Department; 2006

[27] ETSI Standard, Speech processing, transmission and quality aspects (STQ), distributed speech recognition, front-end feature extraction algorithm, compression algorithm, ETSI Standard ES 201 108 v1.1. Valbonne, France; 2000