INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 3, PP. 569–574
Manuscript received June 17, 2024; revised July, 2024. doi: 10.24425/ijet.2024.149580

Amplitude spectrum correction to improve speech signal classification quality

Stanislaw Gmyrek, Robert Hossa, and Ryszard Makowski

***Abstract*—The speech signal can be described by three key elements: the excitation signal, the impulse response of the vocal tract, and a system that represents the impact of speech production through human lips. The primary carrier of semantic content in speech is primarily influenced by the characteristics of the vocal tract. Nonetheless, when it comes to parameterization coefficients, the irregular periodicity of the glottal excitation is a significant factor that leads to notable variations in the values of the feature vectors, resulting in disruptions in the amplitude spectrum with the appearance of ripples. In this study, a method is suggested to mitigate this phenomenon. To achieve this goal, inverse filtering was used to estimate the excitation and transfer functions of the vocal tract. Subsequently, using the derived parameterisation coefficients, statistical models for individual Polish phonemes were established as mixtures of Gaussian distributions. The impact of these corrections on the classification accuracy of Polish vowels was then investigated. The proposed modification of the parameterisation method fulfils the expectations, the scatter of feature vector values was reduced.**

***Keywords*—automatic speech recognition, robust parameterization, amplitude spectrum correction, inverse filtering**

I. Introduction

THERE is a need in automatic speech recognition (ASR) systems to compensate for the influence of many factors such as recording conditions, interpersonal differences, contextuality, etc., which adversely affect the performance of the system. One group of such methods is robust parameterization, which should make the parameter vector resistant to diversity in the aforementioned factors, or at least reduce their impact.

A widely accepted model of speech production is of the form:

$$s(n) = x(n) \star h(n) \star r(n) \tag{1}$$

where $x(n)$ is the excitation, $h(n)$ is the impulse response of the vocal tract, $r(n)$ is the impulse response characterizing the sound emission by the lips, n is the discrete time, and $\star$ is the convolution operator [1]. The semantic information contained in speech is mainly shaped by the vocal tract. On the other hand, the quasiperiodicity of glottal excitation is one of the factors contributing to the significant scatter in the values of their resulting coefficients, by introducing ripples into the amplitude spectrum (see Section II).

S. Gmyrek, R. Hossa, R. Makowski are with Department of Acoustics, Multimedia and Signal Processing, Wroclaw University of Science and Technology, Wroclaw, Poland (e-mail: stanislaw.gmyrek@pwr.edu.pl, robert.hossa@pwr.edu.pl, ryszard.makowski@pwr.edu.pl).

This paper presents a method to mitigate the impact of glottal excitation through a filtering process. First, the excitation signal $x(n)$ is estimated, and then the basis for determining the Human Factor Cepstral Coefficients (HFCC) is the magnitude of the vocal tract transfer function. The estimation of excitation is achieved through known inverse filtering algorithms, which involve removing the effects of components $h(n)$ and $r(n)$ based on their parametric models determined through Linear Predictive Coding analysis (LPC). Ensuring a reliable vocal tract model is crucial in this approach, and there are various methods to achieve this [2] [3] [4] [5]. Notably, options include (i) the Closed Phase Inverse Filtering (CPIF) algorithm, which focuses on analyzing only the closing phase of the vocal cord vibration cycle, and (ii) iterative approaches and synchronization mechanisms like Iterative Adaptive Inverse Filtering (IAIF) and Pitch Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) [6]. In addition to inverse filtering, there are parametric techniques and algorithms that use the mixed-phase model of the speech signal [7]. In this approach, the IAIF algorithm was employed. To evaluate the performance of the proposed parameterization methods, statistical models for individual phonemes in Polish speech were developed using a mixture of Gaussian distributions (GMM model). The purpose of the considered corrections was to narrow the GMM distributions of the amplitude spectrum and simultaneously increasing the distance between them [8]. In general, according to detection theory, it minimizes the classification errors. The assessment of the proposed correction effectiveness was carried out by comparing Frame Error Rate (FER) measurements before and after execution of the correction algorithm. [9].

II. Theory

A. Signal parametrization

From the numerous parameterization techniques found in the literature, the approaches utilizing time-frequency transforms and cepstral representations are recognized as some of the most extensively employed and efficient methods [10] [11]. These include Mel Frequency Cepstral Coefficients (MFCC), Human Factor Cepstral Coefficients (HFCC), Basiliar-membrane Frequency-band Cepstral Coefficient (BFCC), and Gammatone Cepstral Coefficient (GTCC). In this study, the HFCC representation was selected. This method is particularly valuable when working with noisy or adverse acoustic conditions and has found applications in areas

such as speech and speaker recognition, speech synthesis and acoustic scene analysis [12] [13]. The parameterization results in the cepstral coefficient vectors $c(t,m)$, that is

$$c(t,m) = \sum_{j=1}^{J} Y_l(t,j) cos\left(m\left(j-\frac{1}{2}\right)\frac{\pi}{J}\right);\ m = 1,...,M \tag{2}$$

where $Y_l(t,j)$ is the logarithm of the signal spectrum $Y(t,j)$, expressed in Equivalent Rectangular Bandwidth (ERB) scale and obtained from the amplitude spectrum $S(t,f)$ under correction, t is the frame number, j is the ERB-scaled frequency band number, J is the number of frequency bands, and M is the number of HFCC coefficients. Moreover HFCC parametrization scheme utilizes bank of uniformly distributed ERB-scale triangular filters which is designed to mimic the non-linear frequency perception of the human auditory system. It groups the spectral energy into frequency bands to reflect human hearing characteristics. The logarithm of the energy within each frequency band is taken to replicate the logarithmic perception of loudness by the human auditory system. The Human Factor Cepstral Coefficients approach to speech features extraction has been proposed and described in details in [14].

B. The influence of fundamental frequency on HFCC coefficients

For reasons of illustration Fig. 1 shows the amplitude spectra of consecutive frames of phoneme "a" selected from longer utterances by the same speaker, recorded under identical conditions, differing in fundamental frequencies f_0. The key distinction among these spectral representations is the various locations of the local maxima, which are multiples of the frequency f_0. Due to the presence of ripples, the formants are not clearly visible, although their frequencies are approximately: 800 Hz, 1.3 kHz, 2.4 kHz, and 4.0 kHz. In these figures, filters with center frequencies corresponding to the mel scale (as in the HFCC parameterization) are also indicated by dotted lines.

The consequence of the different positions of the local maxima of the spectrum is the different energy per successive frequency band, which leads to different ERB-scale spectra at different f_0. This is confirmed by the plots of ERB-scale spectra presented in Fig. 2. Particularly large differences occur for band 4. As a consequence, there are significant variations in the cepstral coefficients for the two considered cases presented in Fig. 3

C. Spectrum correction

Theoretically, the excitation signal, for each voiced frame, can be determined using inverse filtering procedure [15] [16], i.e.

$$x(n) = s(n) \star (h(n) \star r(n))^{-1}, \tag{3}$$

where $(.)^{-1}$ denotes the inverse in the convolution sense. Introducing $w(n) = x(n) \star r(n)$, i.e. as the convolution of the excitation signal and the function describing the lips radiation, the quantity $w(n)$ can be determined from the relation

$$\tilde{w}(n) = s(n) \star \tilde{h}(n)^{-1}. \tag{4}$$

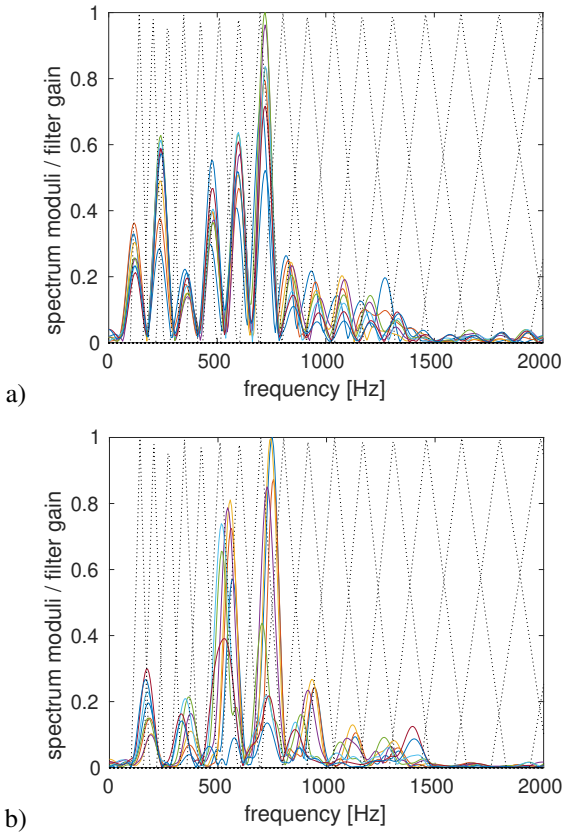

Fig. 1. Amplitude spectra $S(t,f)$ of consecutive frames of phoneme a with applied filterbank (dotted line); the fundamental frequency a) about 130Hz b) about 195Hz.

The relation (4) describes the problem of blind deconvolution. It requires the estimation of the impulse response $h(n)$ and then the determination of the inverse in the convolution sense of this quantity. In general, in this situation, the problem of stability arises. Fortunately a stability condition is guaranteed if the impulse response $h(n)$ is minimum-phase or an algorithm, which enforces minimum phase property, is taken into account in experimental studies. The most popular solution in this case is mean-square filtering [1] which is used in the applied IAIF filtering.

The IAIF block diagram, slightly modified for the purposes of the work, is presented in Fig.4. In the first step, a preliminary estimator of the filter is determined that models the combination of glottal excitation and lip radiation using an LPC filter of order m_1. In the second step, after compensating for the influence of $G_1(z)$ on the signal $s(n)$, a preliminary estimator $H_{v1}(z)$ of the vocal tract is determined with LPC filter of order m_2. The resulting estimator $H_{v1}(z)$, in Step 3, is used to filter out the influence of the vocal tract from the signal $s(n)$. In this step, the influence of the lip emission properties is also eliminated by integration, and a more accurate parametric model $G_2(z)$ is calculated with the LPC filter of order m_1. In the fourth step, using $G_2(z)$, by means of inverse filtering, integration, and LPC analysis, the parameters of the $H_{v2}(z)$ model of the vocal tract of order m_3 are determined. The result of the last operation is used to calculate the HFCC coefficients

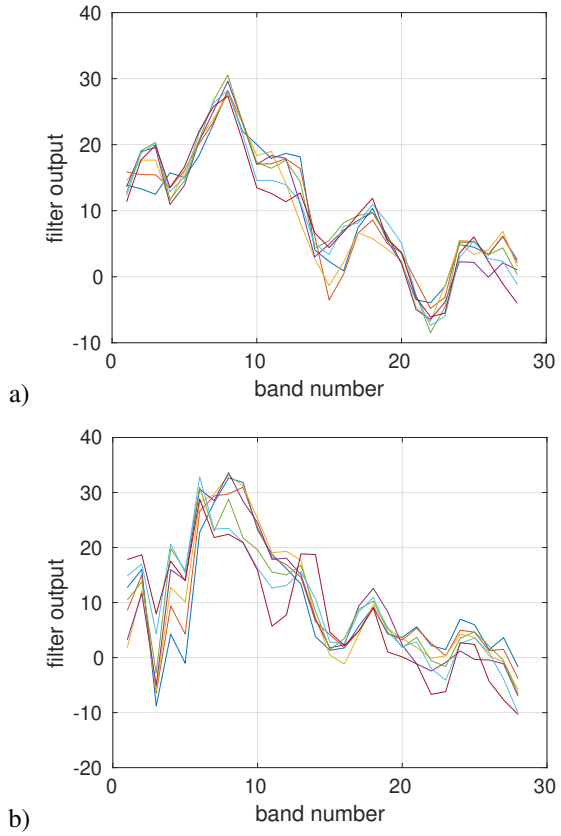

Fig. 2. Spectra of consecutive ERB-scale frames of phoneme a; the fundamental frequency is a) about 130Hz, b) about 195Hz.

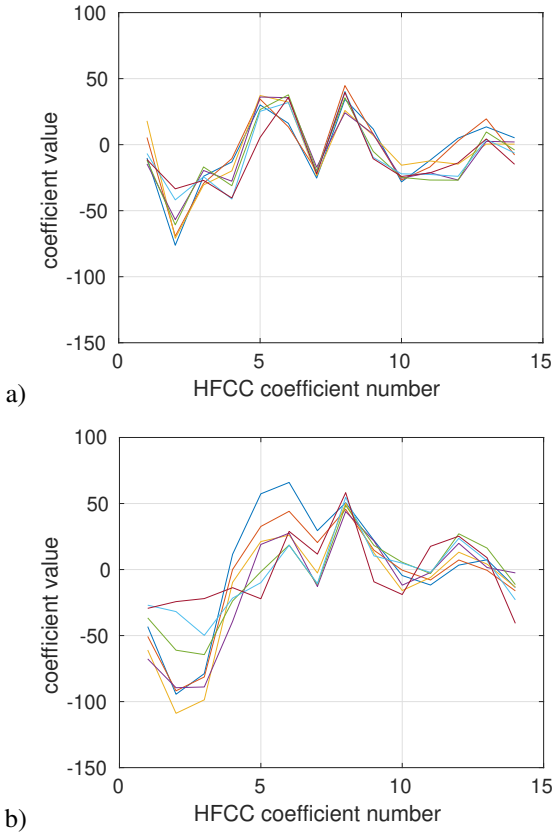

Fig. 3. Cepstra of consecutive ERB-scale frames of phoneme a; the fundamental frequency is a) about 130Hz, b) about 195Hz.

after compensating for the influence of glottal excitation. This method is called cHFCC for the purpose of this paper.

Two facts are worth keeping in mind:

- the results are obtained under the assumption of minimum phase property of all elements of the relation (1),
- since the phase of the signal is not taken into account in the HFCC parameterization, it is hoped that modeling the elements of the relations (1) using the LPC model, will yield the expected results.

D. Correction quality measure

In order to assess the performance of the proposed methods of modifying the parameterization of HFCC, a study was carried out on Polish speech vowels occurring in the section III. The implementation of the concept introduced above required the prior development of acoustic models of these vowels in the form of GMM probability distributions. Moreover the single frame recognition error measure was used to evaluate the effectiveness of compensation.

The GMM acoustic models used at the frame recognition stage are a mixture of K=7 multivariate normal probability distributions with a diagonal covariance matrix $\mathbf{\Sigma}$ determined based on the Expectation-Maximization (EM) algorithm:

$$p_f(\mathbf{o}) = \sum_{i=1}^{K} w_{fi}\mathcal{N}(\mathbf{o}, \mathbf{m}_{f,i}, \mathbf{\Sigma}_{f,i}), \tag{5}$$

where w_{fi}, $\mathbf{m}_{f,i}$ denotes the mixture i^{th} component weights and means for f^{th} phoneme. The EM algorithm was described in detail in [17].

Frame Error Rate (FER) is typically used to evaluate the quality of speech recognition at the individual frame level and is defined as

$$FER = \frac{T_{err}}{T} \cdot 100\% \tag{6}$$

where T is the number of all frames to be recognized and T_{err} is the number of frames incorrectly recognized.

III. Experiments

The set of recordings constituting the database for the experiments consists of 36 adult male voices recorded in different Polish cities. For each speaker, 150 words of Polish were recorded, of which speech fragments containing vowels from 43 words were used in the experiment. The sampling rate of the signals was 12 kHz. The results presented here are for noisy signals with a signal-to-noise ratio of 35 dB. All these recordings were manually segmented and labeled, and the phonetic unit in the labeling is the phoneme. The frame length was 30 ms and the pitch 10 ms. The number of cepstral coefficients was $N = 14$. The speakers were divided into groups, and the criterion for division was based on the cepstral coefficients of the vowels. The division method using the universal backgraund model is described in the paper [18]

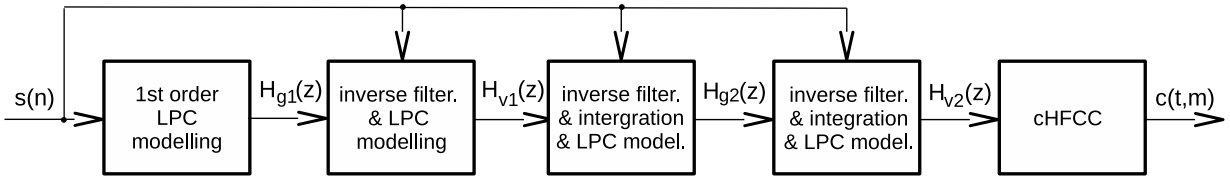

Fig. 4. Block diagram of the applied inverse filtering algorithm (IAIF).

A. *Examplary results*

The section presents example results of the cHFCC method for three consecutive frames of the *a* phoneme, whose statistics are presented in Figs. 1-3. Fig. 5 presents successively (a) amplitude spectra of the signal frames, (b) moduli of coarse estimators $G_1(f)$, (c) transmittance moduli of coarse estimators $H_{v1}(f)$, (d) moduli of estimators $G_2(f)$ and (e) transmittance moduli of estimators $H_{v2}(f)$.

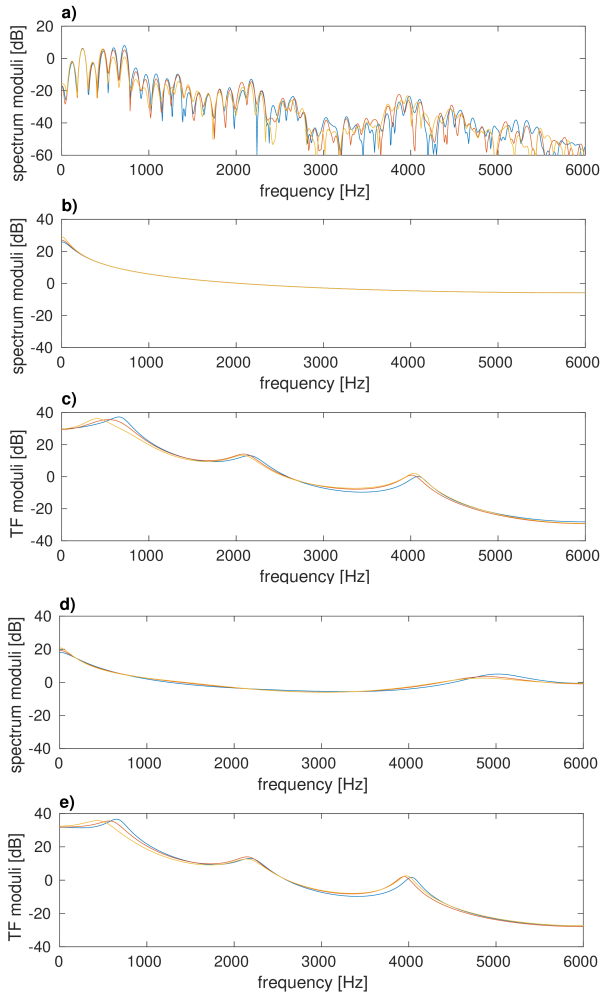

Fig. 5. Example results of the cHFCC method for 3 consecutive frames of phoneme *a*: (a) amplitude spectra of the signal frames, (b) moduli of coarse estimators $G_1(f)$, (c) transmittance moduli of coarse estimators $H_{v1}(f)$, (d) moduli of estimators $G_2(f)$ and (e) transmittance moduli of estimators $H_{v2}(f)$.

Cepstral coefficients were calculated based on the results, examples of which are presented in Fig. 4e). Comparison of graphs (a) with graphs (e) shows the effectiveness of the proposed method for eliminating ripples caused by the quasi-periodicity of glottal excitation.

B. *The impact of selected parameters on correction efficiency*

The efficiency of the presented algorithm is affected by the values of the processing parameters. One of them is the order of the LPC estimation filter H_{g1}-parameter m_1. The classical preemphasis is an order-1 model, but an order-2 model is also used. Fig. 6 presents FER plots for vowels for $m_1 = 1$ and $m_1 = 2$.

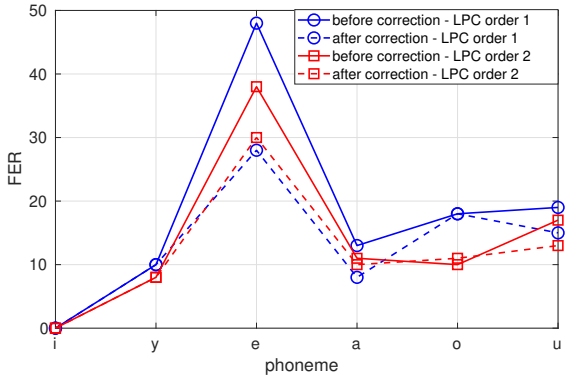

Fig. 6. FER for one of the speakers. $m_2 = 10$, $m_3 = 8$, m_1 changes, diagonal matrix $Sigma_{fi}$, K=7

Fig. 6 shows that a value of $m_1 = 1$ results in smaller FER values after spectrum correction. The effectiveness of spectrum correction can be expected to vary for different speakers in a group. Fig. 7 presents FER plots for the same values of processing parameters for 3 speakers of the same group, before and after correction.

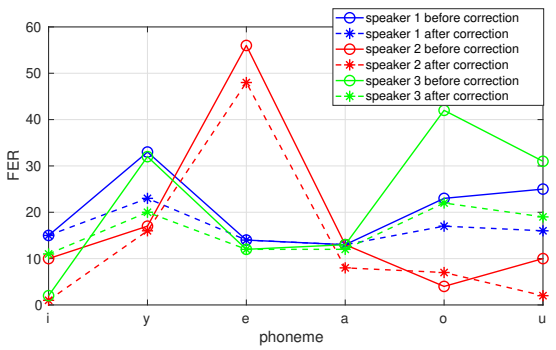

Fig. 7. Global FER values for Polish speech vowels. $m_1 = 1, m_2 = 10, m_3 = 8$, diagonal matrices Σ_{fi}, K=7

The next parameter to consider is the number of elements of the mixture of normal distributions from which the acoustic models of phonemes are built - the number K in the formula 5. In fig. 8 presents FER plots for Polish speech vowels, before and after correction, for values of $K = 7$ and $K = 10$.

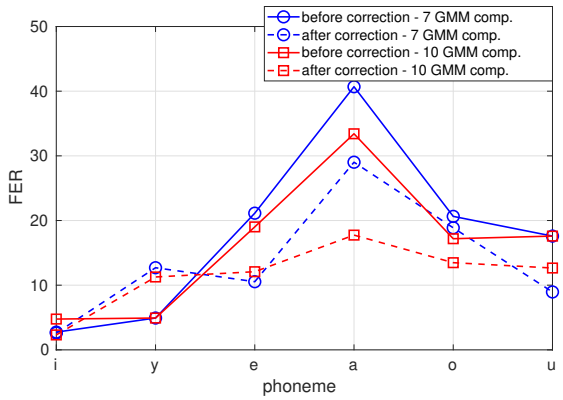

Fig. 8. Global FER values for exemplary speaker. Parameters: $m_1 = 1, m_2 = 10, m_3 = 8$, diagonal matrices Σ_{fi}, K changes

The graphs in Fig. 8 show that FER errors take smaller values for K=10. However, this is not favorable for reasons of computational efficiency. A larger number of observations is also required, which is not always possible, especially for consonants.

C. Global error analysis

In Fig. 9 the results of the FER measure in one-to-one recognition for Polish speech vowels are presented in the form of a table. The upper values indicate the FER before correction and the lower values after correction. The selective color indicates situations for which there was a decrease in FER and the red color indicates an increase.

	i	y	e	a	o	u
i		1.22 0.62	0.95 0.95	0.00 0.00	0.00 0.04	1.69 1.40
y	5.10 4.11		8.35 4.07	1.15 0.00	1.70 0.90	2.32 0.29
e	3.64 1.72	15.07 14.05		3.72 3.26	1.85 1.66	1.79 1.53
a	0.84 0.69	2.12 2.02	7.87 8.32		4.69 4.48	1.19 0.78
o	1.06 0.67	1.64 1.67	3.23 2.93	6.74 7.56		4.43 4.07
u	1.52 1.52	2.19 2.67	2.63 0.52	0.00 0.00	3.34 4.53	

Fig. 9. FER values for Polish speech vowels in one-to-one recognition. The upper/lower values denotes FER before/after correction. Green color indicates the error decrease, red color - the increase.

The results presented in Fig. 9 show that in most cases there was a reduction in the recognition errors of single frames.

The average values of FER for 2 groups of speakers (as the sum of errors in one-to-one recognition) were presented in Fig. 10-11 The group size of Fig. 10 is 6 and that of Fig. 11 is equal to 10. The values of the spectrum correction parameters are as follows: the order of the LPC filter of excitation spectrum estimation m_1=1, the order of the coarse vocal tract transfer function estimation m_2= 10, the order of the finer filter of the vocal tract transfer function estimation m_3= 8, the covariance matrices are diagonal, the number of GMM mixture elements K=7.

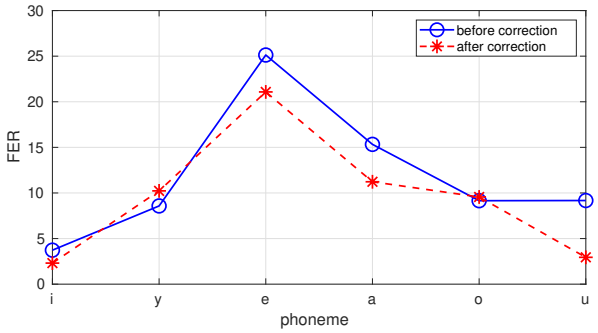

Fig. 10. Global FER values for two groups of Polish speakers. The group size is 6 speakers.

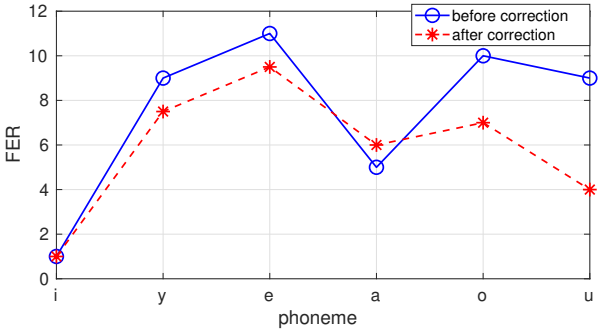

Fig. 11. Global FER values for two groups of Polish speakers. The group size is 10 speakers.

For the majority of Polish speech vowels, spectrum correction allows a reduction in FER values. These are not usually large changes, but in speech recognition systems with high complexity, any improvement in recognition quality is important and desirable.

Another experiment carried out was to compare the performance of speech signal frame recognition for GMM models composed of normal distributions with diagonal covariance matrix or full covariance matrix. The results of this experiment are presented in Fig. 12. In general, the use of full covariance matrices results in smaller frame recognition errors. However, this is disadvantageous due to a much more difficult computational process and higher computational complexity.

IV. Conclusion

The proposed in this paper modification of the HFCC parameterization fulfills the expectations. Through estimation

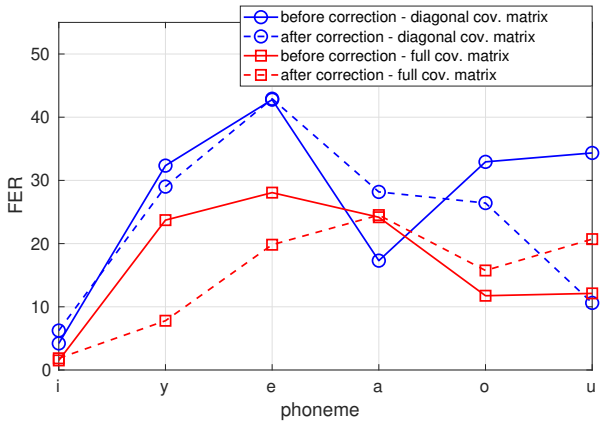

Fig. 12. Global FER values for Polish speech vowels. Comparison of the results of tests conducted using the diagonal and full covariance matrix.

and inverse filtering, it is possible to realize the minimization of the influence of the quasi-periodicity of glottal excitation on the determination of the HFCC coefficients. Consequently, the scatter of feature vector values is reduced. This conclusion is confirmed by the results obtained in experimental studies based on the classification errors of individual frames. As a result, such a modification of the HFCC parameterization should result in an increase of the efficiency of the complete ASR system. At the same time, it should be kept in mind that the variability of the components of the feature vector, in addition to the influence of the quasi-periodicity of the glottal tone, is affected by a number of other factors such as:

- interpersonal variability
- intrapersonal variability
- contextual variability
- the influence of recording condition etc.

REFERENCES

[1] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 1st ed. Upper Saddle River, NJ: Prentice Hall, Oct. 2001.

[2] J. Walker and P. Murphy, *A Review of Glottal Waveform Analysis*, Jan. 2005, vol. 4391, pages: 21. [Online]. Available: https://doi.org/10.1007/978-3-540-71505-4_1

[3] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech Language*, vol. 26, pp. 20–34, 01 2012. [Online]. Available: https://doi.org/10.1016/j.csl.2011.03.003

[4] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, Aug. 1979, conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing. [Online]. Available: https://doi.org/10.1109/TASSP.1979.1163260

[5] M. Plumpe, T. Quatieri, and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999. [Online]. Available: https://doi.org/10.1109/89.784109

[6] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2, pp. 109–118, Jun. 1992. [Online]. Available: https://doi.org/10.1016/0167-6393(92)90005-R

[7] K. Syed and T. Qureshi, "A New Approach to Parametric Modeling of Glottal Flow," *Archives of Acoustics; 2011; vol. 36; No 4; 695-712*, 2011, publisher: Committee on Acoustics PAS, PAS Institute of Fundamental Technological Research, Polish Acoustical Society.

[8] J. Goldberger and H. Aronowitz, "A distance measure between gmms based on the unscented transform and its application to speaker recognition," 09 2005, pp. 1985–1988. [Online]. Available: https://doi.org/10.21437/Interspeech.2005-624

[9] R. Makowski, *Automatyczne rozpoznawanie mowy: wybrane zagadnienia*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2011. [Online]. Available: https://books.google.pl/books?id=qv5vMwEACAAJ

[10] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Communication*, vol. 53, no. 5, pp. 707–715, 2011, perceptual and Statistical Audition. [Online]. Available: https://doi.org/10.1016/j.specom.2010.04.008

[11] T.-W. Kuan, A.-C. Tsai, P.-H. Sung, J.-F. Wang, and H.-S. Kuo, "A robust bfcc feature extraction for asr system," *Artificial Intelligence Research*, vol. 5, 01 2016. [Online]. Available: https://doi.org/10.5430/air.v5n2p14

[12] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[13] M. Skowronski and J. Harris, "Improving the filter bank of a classic speech feature extraction algorithm," in *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.*, vol. 4, 2003, pp. IV–IV. [Online]. Available: https://doi.org/10.1109/ISCAS.2003.1205828

[14] ——, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 116, pp. 1774–80, 10 2004. [Online]. Available: https://doi.org/10.1121/1.1777872

[15] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, Jan. 2011, conference Name: IEEE Transactions on Audio, Speech, and Language Processing. [Online]. Available: https://doi.org/10.1109/TASL.2010.2045239

[16] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–1332, 02 2004. [Online]. Available: https://doi.org/10.1121/1.1646401

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/2984875

[18] R. Hossa and R. Makowski, "An effective speaker clustering method using ubm and ultra-short training utterances," *Archives of Acoustics*, vol. 41, 03 2016. [Online]. Available: https://doi.org/10.1515/aoa-2016-0011