Archives of Acoustics Vol. 50, No. 1, pp. 59–67 (2025), https://doi.org/10.24425/aoa.2025.153652

Research Paper

The Influence of the Amplitude Spectrum Correction in the HFCC Parametrization on the Quality of Speech Signal Frame Classification

Stanislaw GMYREK*, Robert HOSSA, Ryszard MAKOWSKI

Faculty of Electronics, Photonics and Microsystems, Department of Acoustics, Multimedia and Signal Processing
Wrocław University of Science and Technology
Wrocław, Poland; e-mails: robert.hossa@pwr.edu.pl, ryszard.makowski@pwr.edu.pl

*Corresponding Author e-mail: stanislaw.gmyrek@pwr.edu.pl

(*received July 8, 2022; accepted December 11, 2024; published online February 28, 2025*)

The voiced parts of the speech signal are shaped by glottal pulse excitation, the vocal tract, and the speaker's lips. Semantic information contained in speech is shaped mainly by the vocal tract. Unfortunately, the quasiperiodicity of the glottal excitation, in the case of the HFCC parameterization, is one of the factors affecting the significant scatter of the feature vector values by introducing ripples into the amplitude spectrum. This paper proposes a method to reduce the effect of quasiperiodicity of the excitation on the feature vector. For this purpose, blind deconvolution was used to determine the vocal tract transfer function estimator and the corrective function of the amplitude spectrum. Subsequently, on the basis of the obtained HFCC parameters, statistical models of individual Polish speech phonemes were developed in the form of mixtures of Gaussian distributions, and the influence of the correction on the quality of classification of speech frames containing Polish vowels was considered in details. The aim of the introduced solution was to narrow the GMM distributions, which clearly, according to the detection theory, reduces classification errors. The results obtained confirm the effectiveness of the proposed method.

Keywords: automatic speech recognition; robust parametrization; amplitude spectrum correction; inverse filtering; GMM model; distance between GMM distributions.

1. Introduction

In automatic speech recognition (ASR) systems, there is a need to compensate for the influence of many factors, such as recording conditions, interpersonal variability, contextuality, etc., which negatively affect the performance of the system. The most widely used compensation methods are (Makowski, 2011):

1) clustering with developing independent statistical models for speakers with similar personal characteristics (Hossa, Makowski, 2016);
2) normalisation, which involves modifying the values of parametrization coefficients (Prasad, Umesh, 2013);
3) adaptation, involving changing the parameter values of statistical models (Zambrzycka, 2021);
4) robust parametrization (Mrówka, Makowski, 2008), which should make the parameter vector robust to the factors mentioned above or at least reduce their impact.

The present work stands for the robust parametrization.

Among at least a dozen different parametrization methods available in the literature (Sharma *et al.*, 2020), the most commonly used and effective solutions in practical applications include methods that use short time-frequency transformations and cepstral representations of the resulting coefficients. To this group of solutions we can include the algorithms:

- Mel-frequency cepstral coefficients, MFCC (Davis, Mermelstein, 1980);
- human factor cepstral coefficients, HFCC (Skowronski, Harris, 2003);
- the basilar-membrane frequency-band cepstral coefficient, BFCC (Kuan *et al.*, 2016);

– the gammatone cepstral coefficient, GTCC (Yin *et al.*, 2011).

On the other hand, the second group of solutions are algorithms using linear prediction methods and examples of their implementations are the parametrizations:

– linear prediction cepstral coefficients, LPCC (Rabiner Juang, 1993);
– the perceptual linear prediction, PLP (Hermansky, 1990).

Most of the aforementioned parametrizations naturally have mechanisms for robustness against small noise interference, which can be further enhanced by supplementing the method with the relative spectral (RASTA) algorithm to suppress those of the components that are not related to speech articulation. Based on such an idea, the RASTA-PLP hybrid algorithm (Koehler *et al.*, 1994) and the multi-resolution RASTA filtering solution (Hermansky, Fousek, 2005) were developed. Another equivalent representation in the form of the amplitude modulation filter bank (AMFB) has been considered in (Moritz, Kollmeier, 2015). Among the robust parametrization algorithms, we can also distinguish algorithms based on the minimum variance distortionless response (MVDR) the estimator proposed in (Murthi, Rao, 2000) and further developed into the MVDR-MFCC algorithm in (Dharanipragada, Rao, 2001).

In general, the voiced parts of the speech signal are shaped by linear cascade without interactions of the glottal pulse excitation, the vocal tract, and the speaker's lips (Quatieri, 2002). Hence a widely accepted source-filter 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 discrete time convolution operator.

The semantic information contained in speech is mainly shaped by the vocal tract. Unfortunately, the quasiperiodicity of the glottal excitation, in the case of parametrizations based on different time-frequency representations, e.g., MFCC or HFCC, is one of the factors affecting the significant scatter of the feature vector values, by introducing ripples into the amplitude spectrum (see Sec. 2). Furthermore, in (Skowronski, Harris, 2003) it was shown that the HFCC parametrization is characterized by greater robustness to noise than the MFCC and studies have shown differences in recognition performance of up to 30 %. As a result, the classical solution, i.e., the HFCC parametrization, was selected as the representative for further research on ripple reduction.

The paper proposes an algorithm to reduce the impact of glottal flow excitation through its filtering operation. The first step is to estimate the glottal excitation signal $x(n)$ and then determine the HFCC coefficients based on the magnitude of the vocal tract transfer function. The estimation of the excitation signal is one of the most important problems in speech signal processing, and in practical applications it is used, among others, for speaker recognition (Plumpe *et al.*, 1999), analysis of the speaker's emotional state (Waarama *et al.*, 2010) or speech synthesis (Raitio *et al.*, 2011). Inverse filtering algorithms are most commonly used in the literature to filter out the influence of the components $h(n)$ and $r(n)$ of the speech signal model form (Eq. (1)) based on their parametric models determined by the LPC analysis. In this approach, it is important to determine a reliable vocal tract model, which is possible in several ways (Walker, Murphy, 2005). Among them, it is worth mentioning:

1) closed phase inverse filtering, CPIF, the algorithm (Wong *et al.*, 1979) with the closed phase of the vocal cord vibration cycle analysis only;
2) algorithms that use an iterative approach and synchronization mechanisms, e.g., iterative adaptive inverse filtering – IAIF (Alku, 1991; Raitio *et al.*, 2011), and pitch synchronous iterative adaptive inverse filtering – PSIAIF (Alku, 1992).

In addition to inverse filtering, there are also parametric methods (Quereshi, Syed, 2011) and algorithms based on a mixed-phase model of the speech signal. They assume that the impulse response of the vocal tract and the part of the excitation corresponding to the return phase are treated as causal components, while the part of the excitation representing the opening phase in the vocal cord cycle is treated as a non-causal component. Separation of these components can be done using the zeros of the Z-transform (ZZT) algorithm (Bozkurt *et al.*, 2005) or the complex cepstrum decomposition (CCD) algorithm (Drugman *et al.*, 2009). In the present work, as starting point in our research, the IAIF algorithm was used. The elimination of excitation influence are performed for each of the speech frames containing vowels. The HFCC parametrization is then performed, resulting in the cepstral coefficient vectors $c(t,m)$, that is

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

where $Y_l(t,j)$ is the logarithm of the ERB-scaled spectrum $Y(t,j)$ obtained from the amplitude spectrum $S(t,f)$ under correction multiplied by a bank of Mel filters whose widths were determined according to the equivalent rectangular bandwidth (ERB) scale, t is the frame number, j is the Mel band number, J is the number of Mel bands, and M is the number of HFCC coefficients. The use of a Mel scale of frequencies and a nonlinear function on the values of the spectrum allows a better representation of the performance

of the human auditory system by taking into account the nonlinearity of the perception of frequency and intensity of sound. The expected purpose of the amplitude spectrum correction was to narrow the GMM distributions and reduce classification errors. The effectiveness of the proposed solution was evaluated on the basis of the distance between individual GMM distributions and FER measure before and after the correction.

2. The influence of fundamental frequency on HFCC coefficients

Figures 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 (frequency f_0), e.g., for Fig. 1a this is $f_0 \approx 130$ Hz, and for Fig. 1b – $f_0 \approx 195$ Hz.

The main difference between these spectra is in the other positions of the local maxima, which are multiples of the frequency f_0. Furthermore 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 centre frequencies corresponding to the Mel scale (as in the HFCC parametrization) 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 Mel filter band, which leads to different ERB-scaled spectra at different f_0. This can be observed on the plots presented in Fig. 2. Especially large differences are found for the fourth band.

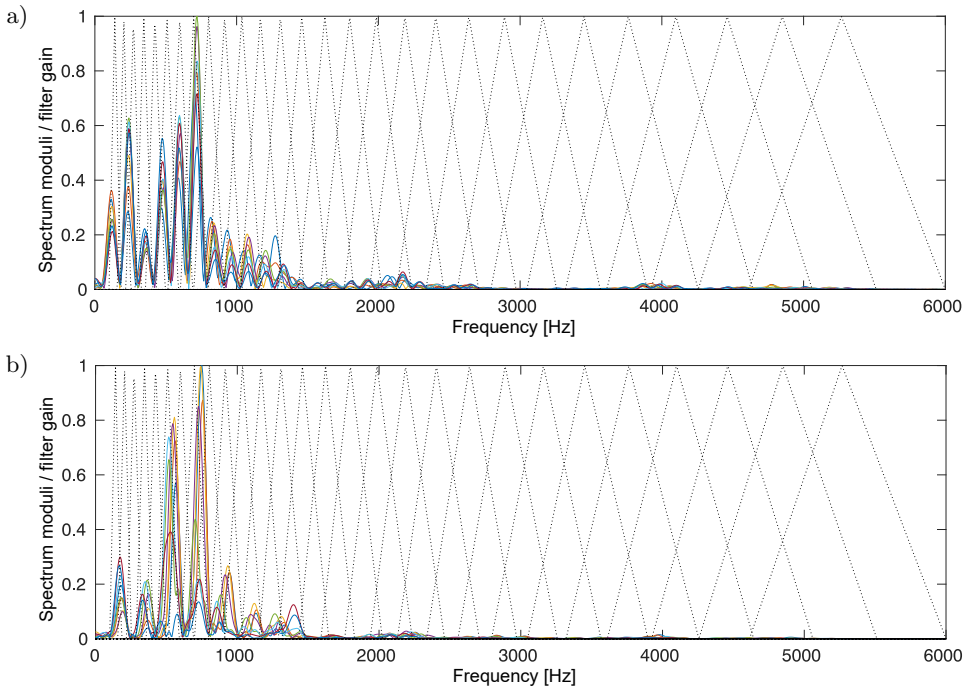

Fig. 1. Amplitude spectra of consecutive frames of phoneme *a* with applied ERB-scale filterbank; the fundamental frequency is about 130 Hz (a) and about 195 Hz (b).

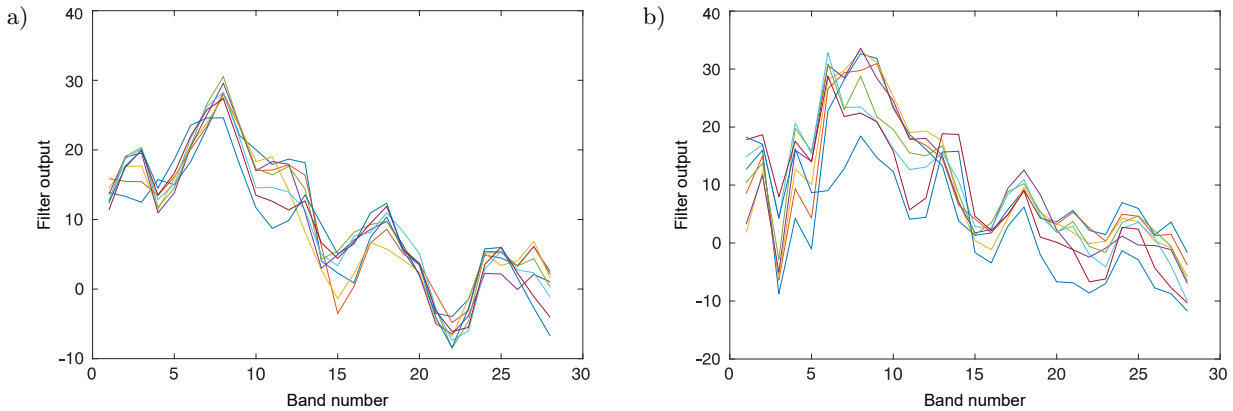

Fig. 2. Spectra of consecutive ERB-scaled frames of the phoneme *a*; the fundamental frequency is about 130 Hz (a) and about 195 Hz (b).

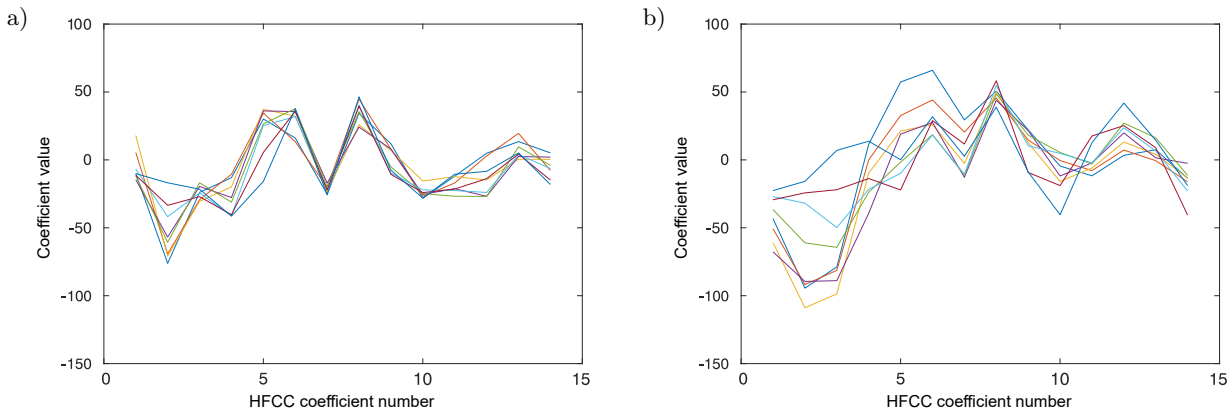

Fig. 3. HFCC coefficients of the phoneme *a* frames; the fundamental frequency is about 130 Hz (a) and about 195 Hz (b).

In turn, Fig. 3 shows plots of HFCC coefficient values for the amplitude spectra presented in Fig. 1. Significant differences can be observed in these figures and the presented examples show the strong influence of the frequency f_0 on the final values of the HFCC coefficients.

3. Glottal excitation signal estimation, correction implementation

In consequence of the experiments analyzed in detail in Sec. 2, the aim of the proposed method is to minimize the effect of excitation signal periodicity on the values of the HFCC coefficients. Theoretically, the excitation signal, for each voiced frame, can be determined using the IAIF (Raitio, 2011; Drugman *et al.*, 2011), i.e.:

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

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 $x(n)$ and the function $r(n)$ describing the lips radiation, the quantity $w(n)$ can be determined from the equation

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

Equation (4) presents a case of the blind deconvolution problem. This operation requires the estimation of the $h(n)$ and then the determination of its inverse in the convolution sense. In the considered situation, the problem of stability can arise, but, this property is guaranteed if the $h(n)$ is a minimum phase or an algorithm, enforcing this minimum phase property, is used. The most popular solution in the case is mean-square filtering (Quatieri, 2002) and is used in the applied pitch synchronized IAIF (PS-IAIF) filtering.

The PS-IAIF block diagram, modified for the purposes of the work, is presented in Fig. 4. In the preprocessing step the estimator YIN (Cheveigné, Kawahara, 2002) for the fundamental frequency f_0 of the input voiced speech is calculated. This algorithm is widely applied in the literature and is known as an effective solution. An input signal $s(n)$ is partitioned, based on the YIN estimator, into frames with length equal to current values of the fundamental period $T_0 = 1/f_0$. Next, for each input frame, in the first step of PS-IAIF, a preliminary estimator of the filter is determined that models the combination of glottal excitation and the lip radiation using an LPC filter of the order the 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 the order 10. The resulting estimator $H_{v1}(z)$, in the third step, 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

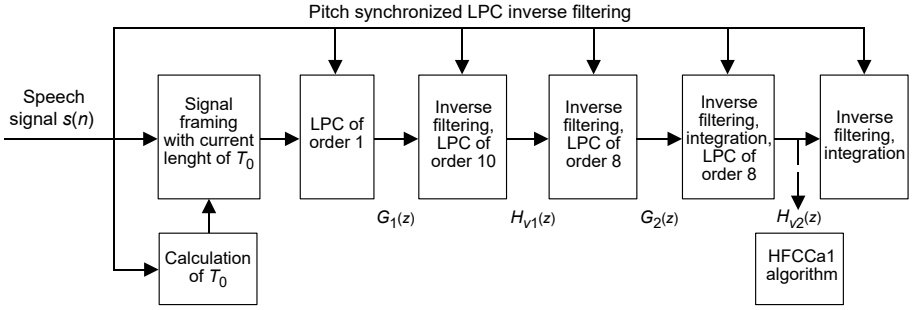

Fig. 4. Block diagram of the applied inverse filtering algorithm (PS-IAIF).

accurate parametric model $G_2(z)$ is determined with the LPC filter of the order 8. 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 the order 8 are determined. Given $H_{v2}(z)$, the frequency domain transfer function is of the form

$$H_{v2}(f) = \frac{1}{1 - \sum_{p=1}^{7} a_p e^{-j2\pi fp/f_s}}. \tag{5}$$

The result of this operation is used to determine the HFCC coefficients after compensating for the influence of the glottal excitation (the HFCCa1 algorithm). Since the phase of the signal spectrum is not taken into account in the HFCC parametrization, we assume here that modelling using the LPC techique will yield minimum phase property of all elements of Eq. (1).

4. Correction quality measures

In order to evaluate the effectiveness of the proposed methods of modifying the HFCC parametrization, numerical tests were carried out on Polish speech vowels occurring in the recording database described in Subsec. 5.1. Performing experiments required the prior development of acoustic models of these vowels in the form of GMM probability distributions, two measures were used to evaluate the effectiveness of the compensation:

1) the Kullback–Leibler distance between the probability distributions (Kullback, 1968) – the KL(·) measure;
2) the single frame error recognition – the FER measure (Makowski, 2011).

4.1. Probabilistic acoustic model of phonemes

The acoustic GMM models used in the frame recognition process are a mixture of $K = 7$ multidimensional normal probability distributions with a diagonal covariance matrices $\mathbf{\Sigma}_{p,i}$ determined based on the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977), i.e.:

$$p_f(\mathbf{o}) = \sum_{i=1}^{K} w_{fi} \mathcal{N}(\mathbf{o}, \mathbf{m}_{f,i}, \mathbf{\Sigma}_{f,i}), \tag{6}$$

where

$$\mathcal{N}(\mathbf{o}, \mathbf{m}_{f,i}, \mathbf{\Sigma}_{f,i}) = \Pi_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{f,i,n}} e^{-\frac{1}{2\sigma^2_{f,i,n}}[o_n - m_{f,i,n}]^2}. \tag{7}$$

4.2. Distances between GMM distributions

In general, a typical measure to calculate the distance between two probability density distributions $p_h(\mathbf{o})$ and $p_g(\mathbf{o})$ for a N-dimensional vector of random variables $\mathbf{o}$ is the Kullback–Leibler divergence defined as follows (Kullback, 1968):

$$\mathrm{KL}(p_h \parallel p_q) = \int_{\mathcal{O}} p_h(\mathbf{o}) \log\left(\frac{p_h(\mathbf{o})}{p_g(\mathbf{o})}\right) \mathrm{d}\mathbf{o}. \tag{8}$$

Unfortunately, for the case of distributions represented by a mixture of Gaussian GMM distributions of the form

$$\begin{aligned} p_h(\mathbf{o}) &= \sum_{i=1}^{K} w_{h,i} \mathcal{N}(\mathbf{o}, \mathbf{m}_{h,i}, \mathbf{\Sigma}_{h,i}) = \sum_{i=1}^{K} w_{h,i} p_{h,i}(\mathbf{o}), \\ p_g(\mathbf{o}) &= \sum_{i=1}^{K} w_{g,i} \mathcal{N}(\mathbf{o}, \mathbf{m}_{g,i}, \mathbf{\Sigma}_{g,i}) = \sum_{i=1}^{K} w_{g,i} p_{g,i}(\mathbf{o}), \end{aligned} \tag{9}$$

where $\mathbf{m}_{h,i}$ and $\mathbf{m}_{g,i}$ are the mean value vectors and $\mathbf{\Sigma}_{h,i}$ and $\mathbf{\Sigma}_{g,i}$ the autocovariance matrices of the components of the Gaussian distributions in the mixtures, there is no closed form formula of the KL(·) measure determination. However, we can use a deterministic approximation of Eq. (8) based on the unscented transform (UT) transformation (Julier, Uhlmann, 2004). Under the assumption that the distributions $p_h(\mathbf{o})$ and $p_g(\mathbf{o})$ are of the GMM form (Eq. (9)) with diagonal covariance matrices, i.e., $\mathbf{\Sigma}_{h,i} = \mathrm{diag}\{\sigma^2_{h,i,k}\}$ and $\mathbf{\Sigma}_{g,i} = \mathrm{diag}\{\sigma^2_{g,i,k}\}$ for $k = 1, 2, ..., N$, we can write that

$$\begin{aligned} \mathrm{KL}(p_h \parallel p_q) &= \int_{\mathbf{O}} p_h(\mathbf{o}) \log\left(\frac{p_h(\mathbf{o})}{p_g(\mathbf{o})}\right) \mathrm{d}\mathbf{o} \\ &= \underset{p_h}{E}[\log p_h(\mathbf{o})] - \underset{p_h}{E}[\log p_g(\mathbf{o})] \\ &= \sum_{i=1}^{K} w_{h,i} \underset{p_{h,i}}{E}[\log p_h(\mathbf{o})] \\ &\quad - \sum_{i=1}^{K} w_{h,i} \underset{p_{h,i}}{E}[\log p_g(\mathbf{o})]. \end{aligned} \tag{10}$$

According to the UT method, for each of the K component distributions of the GMM mixture $p_{h,i}(\mathbf{o}) = \mathcal{N}(\mathbf{o}, \mathbf{m}_{h,i}, \mathbf{\Sigma}_{h,i})$ with diagonal matrices $\mathbf{\Sigma}_{h,i} = \mathrm{diag}\{\sigma^2_{h,i,k}\}$, we generate a set of $2N$ "sigma" points of the form

$$\begin{aligned} \mathbf{o}_{i,k} &= \mathbf{m}_{h,i} - \sqrt{N\sigma^2_{h,i,k}}\mathbf{e}_k, \\ \mathbf{o}_{i,k+N} &= \mathbf{m}_{h,i} + \sqrt{N\sigma^2_{h,i,k}}\mathbf{e}_k, \end{aligned} \tag{11}$$

where $\mathbf{e}_k$ for $k = 1, 2, ..., N$ are basis vectors in the N dimensional Cartesian coordinate system and we determine the approximation of the integral $\underset{p_{h,i}}{E}[\log p_g(\mathbf{o})]$ based on the formula (Goldberger, Aronowitz, 2005)

$$\begin{aligned} \underset{p_{h,i}}{E}[\log p_g(\mathbf{o})] &= \int_{\mathcal{O}} p_{h,i}(\mathbf{o}) \log p_g(\mathbf{o}) \, \mathrm{d}\mathbf{o} \\ &\approx \frac{1}{2N} \sum_{k=1}^{2N} \log p_g(\mathbf{o}_{i,k}). \end{aligned} \tag{12}$$

We include all the partial results of the calculations into Eq. (10) and obtain the approximation of the distance value KL(·) between the considered distributions. To satisfy the symmetry property of the distance measure KL(·) between the GMM distributions $p_g(\mathbf{o})$ and $p_h(\mathbf{o})$, the final form

$$d(p_g, p_h) = \frac{1}{2}(\mathrm{KL}(p_h \parallel p_q) + \mathrm{KL}(p_g \parallel p_h)) \tag{13}$$

was applied in numerical experiments.

4.3. Frame error rate

The frame error rate (FER) is typically used to evaluate the quality of speech recognition at the individual frame level and is defined as

$$m = \frac{T_{\mathrm{err}}}{T} \cdot 100\ \%, \tag{14}$$

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

5. Correction results

5.1. Speech recordings

The set of recordings that constitute the database for the experiments consists of 36 male adult voices recorded in different Polish cities. For each speaker, 150 words of Polish were recorded and speech fragments containing vowels from preliminary chosen 43 words were used in the experiment. The sampling rate of the signals was 12 kHz. The results obtained from numerical experiments are for noisy signals with a signal-to-noise ratio of 35 dB. All of these recordings were manually segmented and labelled, and the phonetic unit in the labelling process was the phoneme. The frame length was 30 ms with the frame shift 10 ms and the number of cepstral coefficients was $N = 14$.

5.2. Examples of algorithm results

The section presents example results of the HFCCa1 algorithm for three consecutive frames of the *a* phoneme, whose statistics are presented in Figs. 1–3. Figure 5 presents successively:

a) the magnitude of the preliminary estimator $G_1(f)$;
b) the magnitude of the transfer function of the preliminary estimator $H_{v1}(f)$;
c) the magnitude of the estimator $G_2(f)$;
d) the magnitude of a transfer function of the estimator $H_{v2}(f)$;
e) the amplitude spectra of the signal frames;
f) the amplitude spectra of the frames after correction.

The cepstral coefficients in the HFCCa1 method are calculated based on the results, examples of which are presented in Fig. 5d. Comparison of plots from Figs. 5d and 5e shows the effectiveness of the proposed algorithms to eliminate ripples caused by the quasiperiodicity of the glottal excitation.

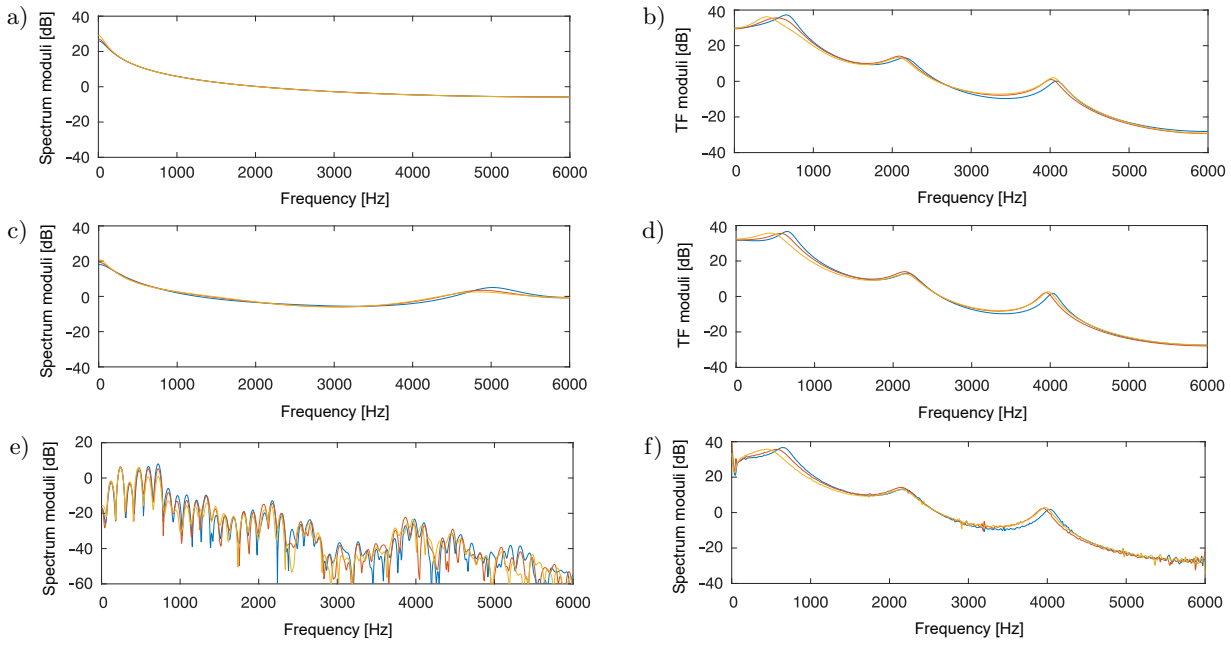

Fig. 5. Example results of the HFCCa1 algorithm for three consecutive frames of the phoneme *a*: a) moduli of the preliminary estimator $G_1(f)$; b) transfer function moduli of the preliminary estimator $H_{v1}(f)$; c) moduli of the estimator $G_2(f)$; d) transfer function moduli of the estimator $H_{v2}(f)$; e) amplitude spectra of the signal frames; f) amplitude spectra of the frames after correction.

5.3. Global results of compensation quality assessment

In Fig. 6, in the form of a table, the KL differences after and before correction between the GMM distributions of the six Polish vowels are presented. Furthermore, the red colour indicates a decrease in the distance after correction and the green colour an increase.

	i	y	e	a	o	u
i		13.46	15.98	1.94	2.931	26.68
y	13.56		19.65	–1.44	3.52	0.47
e	15.98	19.65		0.88	5.89	2.11
a	1.94	–1.44	0.89		23.45	13.69
o	2.31	3.52	5.89	23.45		11.09
u	26.68	0.47	2.11	13.69	11.09	

Fig. 6. Differences in KLD distances after and before correction between the six vowels of Polish speech. The red colour indicates a decrease in distance and the green colour an increase.

It is easily observed that in most cases of comparisons an increase in these distances is observed, and the differences are largest for the phonemes i and u. Simultaneously, significant decreases in distance are noticed between the phonemes y and a. Presenting the results more synthetically, by summing the distances between a given GMM distribution and the other distributions, i.e., determining the values

$$D_f = \sum_{i=1}^{F} d(p_f, p_i) \tag{15}$$

we obtain global KLD distances for individual phonemes before and after correction. These measures are presented in Fig. 7, where it can be seen that an increase in KLD distances occurred for all vowels.

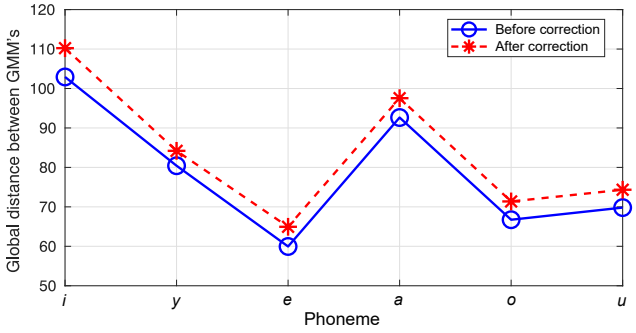

Fig. 7. Global KLD distances for vowels.

In turn, the results of the FER measure in one-to-one recognition for Polish speech vowels are presented in the form of a table in Fig. 8. The upper values in the table elements indicate the FER before correction and the lower values after correction. Furthermore, the green colour indicates situations for which there was a decrease in FER, and the red colour indicates an increase.

	i	y	e	a	o	u
i		2.24 2.22	0.35 1.40	0.00 0.00	0.15 0.04	1.07 1.66
y	1.16 1.07		9.29 7.22	0.06 0.28	0.00 0.00	0.72 0.66
e	1.02 1.35	14.41 13.03		5.43 3.89	0.63 0.98	0.49 0.39
a	0.00 0.00	0.19 0.19	7.11 6.74		2.94 3.07	0.01 0.00
o	0.00 0.00	0.33 0.66	1.39 1.53	4.35 3.02		4.32 4.32
u	1.12 0.68	0.69 1.21	0.08 0.00	0.28 0.25	4.36 3.07	

Fig. 8. FER values for Polish speech vowels.

The results presented in Fig. 8 imply that in most cases there was a reduction in single frame recognition errors. On the other hand, Fig. 9 shows plots of the FER sum following the table rows of Fig. 8.

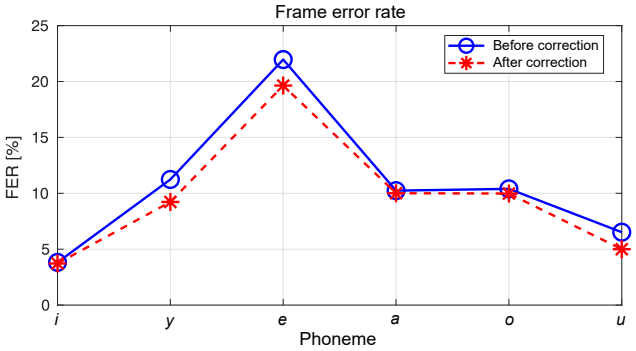

Fig. 9. Global FER values for Polish speech vowels.

This form of obtained data analysis shown in Fig. 9 also confirms that the proposed correction results in a reduction in FER errors.

6. Conclusions

The modification of the HFCC parametrization proposed in this paper meets the predicted expectations. Through estimation and inverse filtering it is possible to minimise the influence of the quasiperiodicity of the source of voiced speech, in the function of the amplitude spectrum $|H_{v_2}(f)|$ used to determine the HFCC coefficients. Consequently, the area of fluctuations of the feature vector values is reduced. This form of the conclusion is confirmed by the obtained results of the Kullback–Leibler distances between the GMM distributions of Polish speech vowels, which are

larger after the correction. Simultaneously, the classification errors of individual frames evaluated by the frame-error-rate measure are also reduced. As a result, the proposed modification of the HFCC parametrization should result in an increase in the efficiency of the complete ASR system. Finally, it should be kept in mind that, in general, the variability of the components of the feature vector, in addition to the considered influence of the quasiperiodicity of the glottal excitation, is affected by a number of other factors such as inter- and intrapersonal variability, contextual variability, influence of recording conditions, etc.

Acknowledgments

Calculations have been carried out using resources provided by the Wrocław Centre for Networking and Supercomputing (grant no. 376).

References

1. Alku P. (1991), Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, [in:] *Proceedings 2nd European Conference on Speech Communication and Technology (Eurospeech 1991)*, pp. 1081–1084, https://doi.org/10.21437/Eurospeech.1991-257.
2. Alku P. (1992), Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, *Speech Communication*, **11**(2–3): 109–118, https://doi.org/10.1016/0167-6393(92)90005-R.
3. Bozkurt B., Doval B., D'Alessandro C., Dutoit T. (2005), Zeros of Z-transform representation with application to source-filter separation in speech, *IEEE Signal Processing Letters*, **12**(4): 344–347, https://doi.org/10.1109/LSP.2005.843770.
4. Cheveigné A., Kawahara H. (2002), YIN, a fundamental frequency estimator for speech and music, *The Journal of the Acoustical Society of America*, **111**(4): 1917–1930, https://doi.org/10.1121/1.1458024.
5. Davis S., Mermelstein P. (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**(4): 357–366, https://doi.org/10.1109/TASSP.1980.1163420.
6. Dempster A.P., Laird N.M., Rubin D.B. (1977), Maximum-likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1): 1–38.
7. Dharanipragada S., Rao B.D. (2001), MCDR based feature extraction for robust speech recognition, [in:] *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 309–312, https://doi.org/10.1109/ICASSP.2001.940829.
8. Drugman T., Bozkurt B., Dutoit T. (2009), Complex cepstrum-based decomposition of speech for glottal source estimation, [in:] *Proceedings of the Annual Conference of the International Speech Communication Association, InterSpeech*, https://doi.org/10.21437/Interspeech.2009-27.
9. Drugman T., Bozkurt B., Dutoit T. (2011), A comparative study of glottal source estimation techniques, *ArXiv*, https://doi.org/10.48550/arXiv.2001.00840.
10. Goldberger J., Aronowitz H. (2005), A distance measure between GMMs based on the unscented transform and its application to speaker recognition, [in:] *9th European Conference on Speech Communication and Technology, InterSpeech*, pp. 1985–1988, https://doi.org/10.21437/Interspeech.2005-624.
11. Hermansky H. (1990), Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, **87**(4): 1738–1752, https://doi.org/10.1121/1.399423.
12. Hermansky H., Fousek P. (2005), Multi-resolution RASTA filtering for TANDEM-based ASR, [in:] *Proceedings of the Annual Conference of the International Speech Communication Association, InterSpeech*, pp. 361–364, https://doi.org/10.21437/Interspeech.2005-184.
13. Hossa R., Makowski R. (2016), An effective speaker clustering method using UBM and ultra-short training utterances, *Archives of Acoustics*, **41**(1): 107–118, https://doi.org/10.1515/aoa-2016-0011.
14. Julier S.J., Uhlmann J.K. (2004), Unscented filtering and nonlinear estimation, [in:] *Proceedings of the IEEE*, **92**(3): 401–422, https://doi.org/10.1109/JPROC.2003.823141.
15. Koehler J., Morgan N., Hermansky H., Hirsch H.G., Tong G. (1994), Integrating RASTA-PLP into speech recognition, [in:] *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, https://doi.org/10.1109/ICASSP.1994.389266.
16. Kuan T.-W., Tsai A.-C., Sung P.-H., Wang J.-F., Kuo H.-S. (2016), A robust BFCC feature extraction for ASR system, *Artificial Intelligence Research*, **5**(2), https://doi.org/10.5430/air.v5n2p14.
17. Kullback S. (1968), *Information Theory and Statistics*, Dover Publications, New York.
18. Makowski R. (2011), *Automatic Speech Recognition – Selected Problems* [in Polish: *Automatyczne Rozpoznawanie Mowy – Wybrane Zagadnienia*], Oficyna Wydawnicza Politechniki Wroclawskiej.
19. Moritz N., Anemuller J., Kollmeier B. (2015), An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(11): 1926–1937, https://doi.org/10.1109/TASLP.2015.2456420.
20. Mrówka P., Makowski R. (2008), Normalization of speaker individual characteristics and compensation of linear transmission distortions in command recognition systems, *Archives of Acoustics*, **33**(2): 221–242.
21. Murthi M.N., Rao B.D. (2000), All-pole modeling of speech based on the minimum variance distortionless response spectrum, *IEEE Transactions on Speech*

and Audio Processing, **8**(3): 221–239, https://doi.org/10.1109/89.841206.

22. Plumpe M.D., Quatieri T.F., Reynolds D.A. (1999), Modeling of the glottal flow derivative waveform with application to speaker identification, *IEEE Transactions on Speech and Audio Processing*, **7**(5): 569–586, https://doi.org/10.1109/89.784109.
23. Prasad N.V, Umesh S. (2013), Improved cepstral mean and variance normalization using Bayesian framework, [in:] *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 156–161, https://doi.org/10.1109/ASRU.2013.6707722.
24. Quatieri T.F. (2002), *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education.
25. Quereshi T.M., Syed K.S. (2011) A new approach to parametric modeling of glottal flow, *Archives of Acoustics*, **36**(4): 695–712, https://doi.org/10.2478/v10168-011-0047-3.
26. Rabiner L., Juang B.-H. (1993), *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs.
27. Raitio T. *et al.* (2011), HMM-Based speech synthesis utilizing glottal inverse filtering, *IEEE Transactions on Audio, Speech and Language Processing*, **19**(1): 1530–165, https://doi.org/10.1109/TASL.2010.2045239.
28. Sharma G., Umapathy K., Krishnan S. (2020), Trends in audio signal feature extraction methods, *Applied Acoustics*, **158**: 107020, https://doi.org/10.1016/j.apacoust.2019.107020.
29. Skowronski M., Harris J.G. (2003) 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*, pp. 281–284, https://doi.org/10.1109/ISCAS.2003.1205828.
30. Waaramaa T., Laukkanen A.M., Airas M., Alku P. (2010), Perception of emotional valences and activity levels from vowel segments of continuous speech, *Journal of Voice*, **24**(1): 8–30, https://doi.org/10.1016/j.jvoice.2008.04.004.
31. Walker J., Murphy P. (2005), A review of glottal waveform analysis [in:] *Progress in Nonlinear Speech Processing, Workshop on Nonlinear Speech Processing, Lecture Notes in Computer Science*.
32. Wong D., Markel J., Gray A. (1979), Least squares glottal inverse filtering from the acoustic speech waveform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**(4): 350–355, https://doi.org/10.1109/TASSP.1979.1163260.
33. Yin H., Hohmann V., Nadeu C. (2011), Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency, *Speech Communication*, **53**(5): 707–715, https://doi.org/10.1016/j.specom.2010.04.008.
34. Zambrzycka A. (2021), *Adaptation in automatic speech recognition systems* [in Polish: *Adaptacja w systemach automatycznego rozpoznawania mowy*], Ph.D. Thesis, Wrocław University of Science and Technology.