INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2025, VOL. 71, NO. 3, PP. 1-7
Manuscript received May 19, 2025; revised July 2025. doi: 10.24425/ijet.2025.153614

# Robust speech parametrization based on pitch synchronized cepstral solutions

Stanisław Gmyrek, and Robert Hossa

***Abstract*—In general, the speech signal can be described by the excitation signal, the impulse response of the vocal tract, and a system that describes the impact of speech emission through human lips. The characteristics of the vocal tract primarily shape the semantic content of speech. Regrettably, the irregular periodicity of glottal excitation represents a significant factor in generating substantial distortions (ripples) in the amplitude spectrum of voiced speech. In this study, a PS-STFT (Pitch-Synchronized Short-Time Fourier Transform) method was proposed to achieve a reliable amplitude spectrum of the vocal tract. Subsequently, a set of cepstral coefficient vectors, namely PS-HFCC (Pitch Synchronized Human Factor Cepstral Coefficients), as a chosen representative of the commonly used classical cepstral parameterization methods was analyzed to investigate the statistical properties after correction. Additionally, the widely accepted in speech recognition applications, the GMM (Gaussian Mixture Model) was chosen as the statistical acoustic model of individual Polish speech phonemes. To evaluate the quality of the proposed method, the distances between the multivariate probability distributions of the GMM form were calculated. Modifying classical cepstral methods through the analysis of variable-length signal frames synchronized to the fundamental period resulted in a reduction in the variance of the estimators of the cepstral coefficients, leading to an increase in the distances between the probability distributions and, consequently, improved classification results.**

***Keywords*—robust cepstral parameterization; fundamental period; amplitude spectrum correction; pitch synchronized STFT**

## I. Introduction

In Speech Processing Systems (SPS), there is a need to compensate for the impact of numerous factors that can negatively affect system performance. These factors include recording conditions, intra- and interpersonal variability, contextuality, etc.

In general, the speech signal can be described by the excitation signal, the impulse response of the vocal tract, and a system that describes the impact of speech emission through human lips. The characteristics of the vocal tract primarily shape the semantic content of speech. Regrettably, the irregular periodicity of glottal excitation represents a significant factor in the generation of substantial distortions (ripples) in the amplitude spectrum of voiced speech.

The paper proposes a new solution, in the general form of Pitch Synchronized Cepstral Parametrization, to significant reduction of this undesirable excitation influence. The first step is to estimate the fundamental period $T_0$ of the input speech signal and then determine the amplitude spectrum using STFT (Short-Time Fourier Transform) with a time-varying window of length consistent with the current value of $T_0$.

A review of the literature reveals at least a dozen different parameterization methods, the most commonly used and effective solutions in practical applications being those that employ time-frequency transformations and cepstral representations. This group of solutions includes the MFCC (Mel Frequency Cepstral Coefficients) [1], HFCC (Human Factor Cepstral Coefficients) [2], BFCC (Basiliar-membrane Frequency-band Cepstral Coefficient) [3], GTCC (Gammatone Cepstral Coefficient) [4] and AMS (Amplitude Modulation Spectrum) [5] algorithms. On the other side there exists another group of solutions using linear prediction methods and examples of their implementations are LPCC (Linear Prediction Cepstral Coefficients) [6] and PLP (Perceptual Linear Prediction) [7] with improvement in the form of the RASTA (Relative Spectra) algorithm [8] or bank of band-pass filters (multi-resolution RASTA filtering) [9].

In [10] 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% [27]. As a result, the HFCC parameterization, was selected as the representative of cepstral parametrization methods for the experimental study on ripple reduction in the amplitude spectrum. The purpose of such an analysis was to check the statistical properties of the classical HFCC and proposed PS-HFCC cepstral coefficient vectors for individual vowels based on the variances of their components. Additionally, the widely accepted in speech recognition applications, the GMM (Gaussian Mixture Model) was chosen as the statistical acoustic model of individual Polish speech phonemes. To evaluate the quality of the proposed method, the distances between the multivariate probability distributions of the GMM form and Frame Error Rate (FER) were calculated.

## II. THEORY

### A. Model of speech signal emission

Commonly accepted in literature and verified experimentally the mathematical model of Fant's source-filter type for a discrete-time speech signal s(n) can be expressed as below [11]:

$$s(n) = v(n) \star l(n) \star u(n) = h(n) \star u(n) \qquad (1)$$

S. Gmyrek, R. Hossa 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)

where $u(n)$ is the excitation, $v(n)$ is the impulse response of the vocal tract filter, $l(n)$ describes the form of speaker speech emission and $\star$ is the discrete convolution operator. The objects of our further considerations are the voiced parts of speech with excitation model $u(n)$ given in the impulse formula [6]:

$$u(n) = g(n) * p(n) = \sum_{k=0}^{+\infty} g(nT_s - kT_0), \tag{2}$$

where $g(n)$ is a single excitation pulse, $p(n) = \sum_{k=0}^{+\infty} \delta(nT_s - kT_0)$ is a pulse train with a repetition time $T_0$ (pitch) while $T_s$ is the sampling interval. In the case of mixed forms of excitation (e.g. plosives) deterministic part of excitation given in the same form (3) can be extracted. In consequence, a voiced part of speech signal $s(n)$ can be written as:

$$s(n) = \sum_{k=0}^{+\infty} s_p(nT_s - kT_0), \tag{3}$$

where $s_p(n)$ is the response of the modeling system to a single input excitation pulse $\delta$(n).

In practical situations, we consider a finite in time representation of the signal, i.e. $s_w(n)$, as the result of the windowing techniques application with the function $w(n)$ on the signal $s(n)$:

$$s_w(n) = s(n)w(n), \tag{4}$$

changing the spectrum S($\omega$) of the signal $s(n)$ to the form [11]:

$$S_w(\omega) = DTFT\{s(n)\} \star DTFT\{w(n)\} = S(\omega) \star W(\omega) \tag{5}$$

where DTFT{·} is the Discrete Time Fourier Transform operator. The influence of spectral leakage introduced with $W(\omega)$ element can be compensated by windowing operation with a properly chosen window function (e.g. Hamming, Kaiser).

### *B. New concept of speech signal spectrum analysis with specific length of the time window*

In general, when analyzing voiced phonemes, a fixed frame length $T_w$ (typically around 30 ms) is assumed, with overlapping between frames (e.g., 30%), while also accounting for possible variations (up to several percent) in the repetition period $T_0$ between neighboring elementary signals $s_p(nT_s)$. Under such conditions, situations often arise where there is a non-integer number of repetitions of the signals $s_p(nT_s)$ within the analysis frame, which makes spectral analysis and the derivation of the correction term formula a highly complex task. For this reason, in further considerations, we simplify the problem and assume that the time length of the analysis window $T_w$ fulfills the condition $T_w = NT_0$, where $N$, under the assumption of the local stationarity of $T_0$ in the analyzed frame, is an integral multiple of the current fundamental period. Even in this highly simplified situation, we obtain the following relation:

$$s_w(n) = w(n) \sum_{k=0}^{N-1} s_p(nT_s - kT_0), \tag{6}$$

with the following spectral representation:

$$S_w(\omega) = \left\{ S_p(\omega) \sum_{k=0}^{N-1} e^{-j\omega kT_0} \right\} * W(\omega) = \tag{7}$$

$$= \left\{ S_p(\omega) \frac{(sin(\omega T_0 \cdot N/2))}{(sin(\omega T_0/2))} e^{-j\omega(N-1)/2T_0} \right\} * W(\omega),$$

which illustrates the source of significant ripples in the amplitude spectrum of the analyzed frame signal $s_w(n)$ what can be observed in Fig 1. Fortunately, for N = 1 the influence of the time-varying $T_0$ on the observed spectrum $S_w(\omega)$ completely disappears, but simultaneously implies the necessity of the current value $T_0$ estimation and suggests the variable-length frames processing.

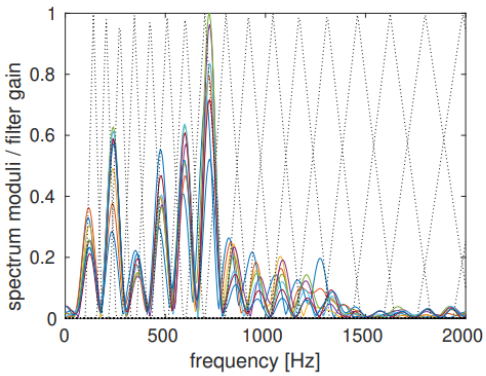


Fig.1. Amplitude spectra $S_w(f)$ of constant length N consecutive frames $s_w(n)$ of exemplary longer utterance of phoneme "a" with the fundamental frequency $f_0$ about 130Hz. The individual amplitude spectra lines of successive frames of the phoneme 'a' have been distinguished using different colors.

The illustration in Fig. 1 shows the amplitude spectra of consecutive frames of Polish phoneme ”a” selected from longer utterances by the same speaker, recorded under identical conditions with fundamental frequencies about $f_0$ = 130 Hz. Due to the existence of ripples of significant levels as local maxima, which are in fact harmonics of the frequency $f_0$, the formants are not easily recognized.

### *C. Fundamental period $T_0$ estimation*

From equation (7) it follows that the linear spectrum compensation procedure requires knowledge of the current $T_0$ value i.e. the use of a simple and efficient method for its estimation. The different solutions for determining pitch provide other effectiveness in noise robustness, accuracy, and computation time. In general solutions to the problem of calculating the current value of $T_0$ use the samples from the analysis frame (e.g. autocorrelation methods [12], [13]) or periodicity in their spectrum (e.g. summation of harmonics algorithms [14], [15]). One of the most popular and efficient methods of pitch estimation is the YIN algorithm [16], together with its statistically improved version [17]. In the numerical experiments of this paper, the YIN solution based on cumulative mean normalized difference function (CMNDF) was chosen and implemented. In general, the computational complexity of the YIN algorithm is $O(W \cdot K)$ where W denotes the length of the analysis window and K represents the maximum lag, expressed

in samples, corresponding to the lowest possible fundamental frequency.

### *D. Cepstral methods of speech parametrization*

Among parameterization solutions widely applied in the literature, the approaches utilizing time-frequency transforms and cepstral representations are recognized as some of the most extensively employed and efficient methods [18]. Generalized form of data flow chart describing cepstral parameterization methods is shown in Fig.2 and is consistent with Mel Frequency Cepstral Coefficients (MFCC), Human Factor Cepstral Coefficients (HFCC), Basiliar-membrane Frequency-band Cepstral Coefficient (BFCC), Gammatone Cepstral Coefficient (GTCC) and Amplitude Modulation Spectrum (AMS) methods. In this paper, the HFCC representation of the input

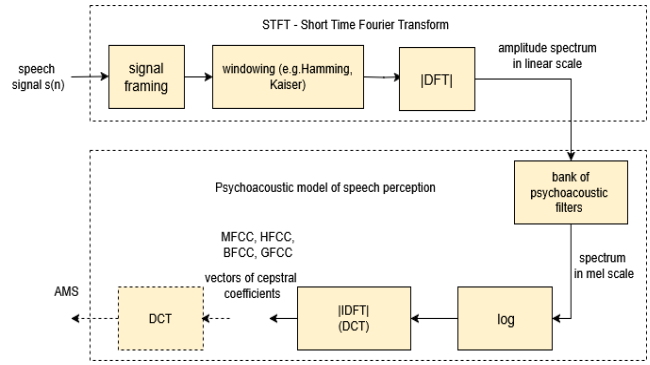


Fig.2. Generalized form of cepstral parameterization methods

frames was taken into consideration. This method is robust to noisy or adverse acoustic conditions and was successfully implemented and verified in speech and speaker recognition, speech synthesis, and acoustic scene analysis [10]. The parameterization results in the cepstral coefficient vectors c(t, m) [1]

$$c(t,m) = \sum_{j=1}^{J} Y_l(t,j) cos\left(m\left(j-\frac{1}{2}\right)\frac{\pi}{J}\right); m = 1,\dots,M \quad (8)$$

where $Y_l(t,j)$ is the amplitude spectrum $S(t,f)$ expressed in mel scale using a bank of filters whose bandwidths have been calculated in the ERB scale, $t$ is the input frame number, $j$ is frequency band number, $J$ is the total number of frequency bands, and $M$ is the number of HFCC coefficients. In this method, the bank of uniformly distributed in ERB-scale triangular filters and the logarithm function implements the perception of the human auditory system. A complete description of the HFCC approach to speech features extraction can be found in [10]. The quasiperiodicity of the glottal excitation and significant fluctuations in the amplitude spectrum produce additional variability in the resulting HFCC coefficients (see Fig.3). The influence of fundamental frequency f0 on HFCC coefficients was considered in details in [19] [20].

### *E. New method of speech parameterization - Pitch synchronized STFT (PS-STFT)*

Relation (7) describing the DTFT spectrum of the analyzed frame

$$S_w(\omega) = \left\{S_p(\omega)\frac{(si\,n(\omega T_0 N/2))}{(si\,n(\omega T_0/2))}\cdot e^{-j\omega(N-1)/2T_0}\right\} \star W(\omega),$$

directly imposes the rules for compensating its distortions introduced with fundamental frequency $f_0$ and their harmonic frequencies. In fact, the choice N = 1 is equivalent to signal analysis with frame of length $T_0$. In this case, the deformation component of desired spectrum $S_p(\omega)$ completely disappears

$$S_w(\omega) = S_p(\omega) \star W(\omega)$$

but introduces the requirement to estimate the current value of $T_0$.

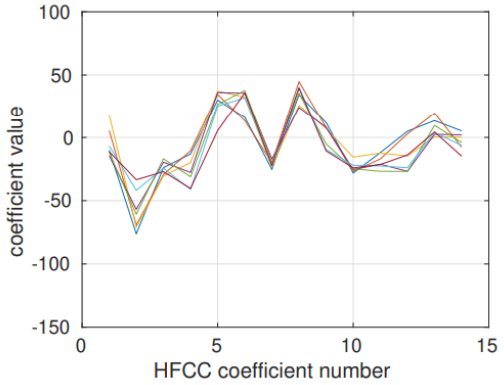


Fig.3. Cepstra of consecutive frames of phoneme “a” with the fundamental frequency about 130Hz.

In the second step the influence of the rectangular window with length $T_0$) on the observed spectrum $S_w(\omega)$ is compensated by using classical solutions of windowing techniques (e.g. Hamming or Kaiser window). Finally, zero padding techniques to obtain a constant length of 1024 sample frame are applied to obtain the same uniform sampling effect of the DTFT transform in the calculated DFT spectrum. This approach also enables the application of classical techniques for averaging the spectra of adjacent frames to compensate for the presence of additive noise (e.g., the Bartlett method). The generic scheme of the proposed PS-STFT method is depicted in Fig. 4.

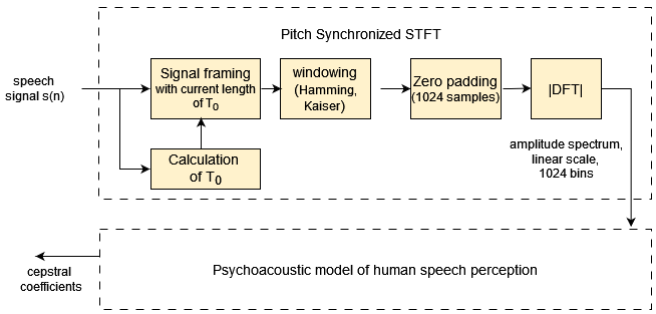


Fig.4. Generalized form of cepstral parameterization methods with the proposed pitch synchronized STFT (PS-STFT)

### F. Gaussian Mixture Model of Polish speech vowels

In order to evaluate the performance of the proposed approach, a study was carried out on Polish speech vowels occurring in section III with their HFCC cepstral parameterization. In effect introduced above cepstral representation required the calculations of vowel acoustic models based on Gaussian Mixture Model (GMM) probability distributions with diagonal covariance matrices form [26]

$$p_f(\boldsymbol{o}) = \sum_{i=1}^{K} w_{fi}\, N\big(\boldsymbol{o}, \boldsymbol{m}_{f,i}, \boldsymbol{\Sigma}_{f,i}\big), \tag{9}$$

where

$$N\big(\boldsymbol{o}, \boldsymbol{m}_{f,i}, \boldsymbol{\Sigma}_{f,i}\big) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{f,i,n}} e^{-\frac{1}{2\sigma_{f,i,n}^2}[o_n - m_{f,i,n}]^2}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{f,i}} = \begin{bmatrix} \sigma_{fi1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{fi2}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{fiM}^2 \end{bmatrix}$$

and $w_{f,i}$, $\boldsymbol{m}_{f,i}$ denotes the mixture $i^{th}$ component weights and means for $f^{th}$ phoneme. GMM acoustic model is usually determined with the EM algorithm ([21]).

### G. Effectiveness measures of pitch synchronized spectrum correction

#### 1) Distance between GMM distributions

In general, a measure to calculate the distance between the 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 [22]

$$KL\big(p_h \parallel p_q\big) = \int_O p_h(\boldsymbol{o}) \log\left(\frac{p_h(\boldsymbol{o})}{p_g(\boldsymbol{o})}\right) d\boldsymbol{o} \tag{10}$$

Unfortunately, for the case of distributions represented by a mixture of Gaussian GMM distributions of the form [26]:

$$p_h(\boldsymbol{o}) = \sum_{i=1}^{K} w_{h,i} N\big(\boldsymbol{o}, \boldsymbol{m}_{h,i}, \boldsymbol{\Sigma}_{h,i}\big) = \sum_{i=1}^{K} w_{h,i} p_{h,i}(\boldsymbol{o})\,;$$

$$p_g(\boldsymbol{o}) = \sum_{i=1}^{K} w_{g,i} N\big(\boldsymbol{o}, \boldsymbol{m}_{g,i}, \boldsymbol{\Sigma}_{g,i}\big) = \sum_{i=1}^{K} w_{g,i} p_{g,i}(\boldsymbol{o}),$$

where $\boldsymbol{m}_{h,i}$ and $\boldsymbol{m}_{g,i}$are the mean value vectors and $\boldsymbol{\Sigma}_{h,i}$ and $\boldsymbol{\Sigma}_{g,i}$the autocovariance matrices of the components of the Gaussian distributions in the mixtures, there is no analytical formula for determining such values. Consequently, we can use approximation methods of stochastic nature in this case, i.e. Monte Carlo simulation methods using the formula [22]

$$KL\big(p_h \parallel p_q\big) = \frac{1}{D}\sum_{i=1}^{D} \log\frac{p_h(\boldsymbol{o}_i)}{p_g(\boldsymbol{o}_i)}, \tag{11}$$

which generally require a huge amount of D data in the form of multidimensional observations oi or their generation based on the known form of the GMM distribution for $p_h(\mathbf{o})$. However, in the lack of such data, we can use a deterministic approximation of the expression (11) using the UT (Unscented Transform) concept [23]. Under the assumption that the distributions $p_h(\mathbf{o})$ and $p_g(\mathbf{o})$ are of the form GMM (9) with K components with diagonal covariance matrices, i.e. $\boldsymbol{\Sigma}_{h,i} = diag\{\sigma_{h,i,k}^2\}$ and $\boldsymbol{\Sigma}_{g,i} = diag\{\sigma_{g,i,k}^2\}$ for $k = 1, 2, \ldots, N$, we can write that [24]

$$KL\big(p_h \parallel p_q\big) = \int_O p_h(\boldsymbol{o}) \log\left(\frac{p_h(\boldsymbol{o})}{p_g(\boldsymbol{o})}\right) d\boldsymbol{o} = \tag{12}$$

$$= \sum_{i=1}^{K} w_{h,i} p_{h,i} E[\log p_h(\boldsymbol{o})] - \sum_{i=1}^{K} w_{h,i} p_{h,i} E\big[\log p_g(\boldsymbol{o})\big]$$

To simplify the description of the approximation procedure of the expression (12), we analyze only the second component of the above sum, since the first one, assuming that $p_g(\mathbf{o}) = p_h(\mathbf{o})$ is its special case. According to the UT method, for each of the K component distributions of the GMM mixture $p_{h,i}(\mathbf{o}) = N\big(\boldsymbol{o}, \boldsymbol{m}_{h,i}, \boldsymbol{\Sigma}_{h,i}\big)$ with diagonal matrices $\boldsymbol{\Sigma}_{h,i} = diag\{\sigma_{h,i,k}^2\}$ we propose a set of $2N$ "sigma" points of the form

$$\boldsymbol{o}_{i,k} = \boldsymbol{m}_{h,i} + \sqrt{N\sigma_{h,i,k}^2}\,\boldsymbol{e}_k;$$

$$\boldsymbol{o}_{i,k+N} = \boldsymbol{m}_{h,i} + \sqrt{N\sigma_{h,i,k}^2}\,\boldsymbol{e}_k,$$

Where $\boldsymbol{e}_k$ for $\mathrm{k} = 1, 2, \ldots, \mathrm{N}$ are basis vectors in the $N$ dimensional Cartesian coordinate system and we determine the approximation of the integral $p_{h,i}E\big[\log p_g(\boldsymbol{o})\big]$ based on the formula [24]

$$p_{h,i}E\big[\log p_g(\boldsymbol{o})\big] \approx \frac{1}{2N}\sum_{k=1}^{2N} \log p_g\big(\boldsymbol{o}_{i,k}\big). \tag{13}$$

We insert all the partial results of the fractional calculation obtained in this way into the relation (12) and obtain the approximation of the distance value between the distributions. To satisfy the symmetry property of the applied distance measure $d$ between the GMM distributions $p_g(\mathbf{o})$ and $p_h(\mathbf{o})$ of the form (12), we take the relation as its final form [22]

$$d\big(p_g, p_h\big) = \frac{1}{2}\Big(KL\big(p_h \parallel p_q\big) + KL\big(p_g \parallel p_h\big)\Big). \tag{14}$$

#### 2) Frame Error Rate

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

$$FER = \frac{T_{err}}{T} \cdot 100\% \tag{15}$$

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

#### 3) Standard deviation of observation vector elements

The last measure of effectiveness selected for comparative analysis of the proposed PS-HFCC method with the classical HFCC approach are standard deviations calculated separately for each coordinate of the cepstral parameter vector. The expected and desired result in this case is the reduction of their

values for each coordinate, which suggests a smaller spread of data and, therefore, a smaller area of their occurrence.

## III. CORRECTION RESULTS

The chapter presents exemplary results of the application of the varying frame length of the speech signal synchronized to the fundamental period $T_0$. In our experiments, the number of mel-bands was 29, cepstral coefficients was $N = 14$. The frame length was 30 ms with the shift 10 ms. The vowel probabilistic acoustic models used in the recognition stage of the research are a mixture of $K = 7$ multidimensional normal probability distributions with diagonal covariance matrices.

### *A. Speech database*

The database for the experiments consists of the recordings of 40 adult male voices recorded in various Polish cities with a sampling rate of 12 kHz and a signal-to-noise ratio of 35 dB. For sampling rate 12 kHz and 1024-point calculated DFT, the frequency resolution $\Delta f$ is 11.71785 Hz. For each mentioned speaker 150 words of Polish were recorded. All these recordings were manually segmented and labeled and finally, the set of more than 100000 of signal pieces with vowels as the phonetic unit was obtained.

### *B. Pitch synchronized STFT experiments*

Exemplary the amplitude spectrum of the phoneme ‘a’ of Polish speech considered in Fig1. and calculated for several consecutive frames with the PS-STFT method is depicted in Fig.5. Comparison of Fig.1 and Fig.5 shows the evident effectiveness of the proposed method for the removal of the amplitude spectrum ripples introduced by the quasi-periodicity of the excitation. Moreover, the four formants of Polish phoneme a′ are clearly visible and their frequency values are in order with commonly accepted tables of their occurrences [25]. This standard form of the spectrum enables the automatic and precise determination of formants and the dynamics of their changes in many different problems related to speech processing.

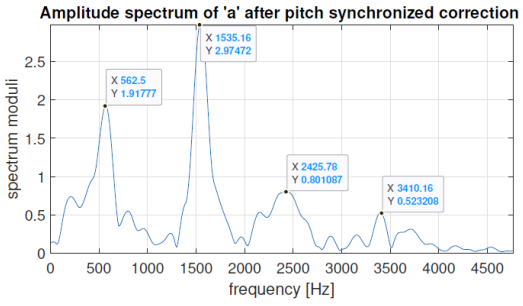


Fig.5. The new form of amplitude spectrum of phoneme “a” considered in Fig.1 as a result of the proposed pitch synchronized STFT with averaging of 3 consecutive frames

### *C. Spectrum correction efficiency*

Consequently, the cepstral representation for each signal frame was calculated using the standard and pitch- synchronized method. In turn, charts depicted in Fig. 6 and Fig. 7 show the standard deviations of the values of the individual cepstral coefficients for the two selected phonemes: the vowel ‘e’ (6) and ‘a’(7). The blue curve represents the standard deviation of the HFCC coefficients, while the red curve represents the standard deviation of the PS-HFCC coefficients. It is evident, that the standard deviations after correction for both analyzed vowels are smaller for each HFCC feature vector coefficient separately which clearly implies a reduction in the area of their occurrence in multidimensional space. Similar observations and results were obtained for all analyzed states, i.e. vowels of Polish speech.

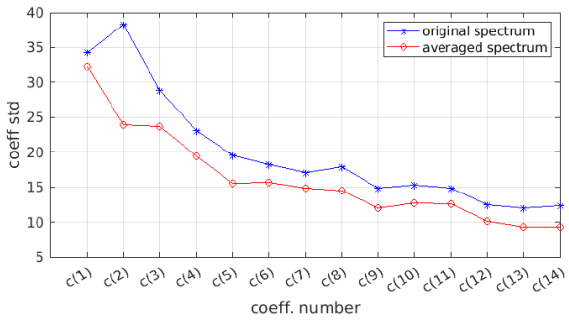


Fig.6. Standard deviations of HFCC cepstral coefficients for the vowel ‘e’

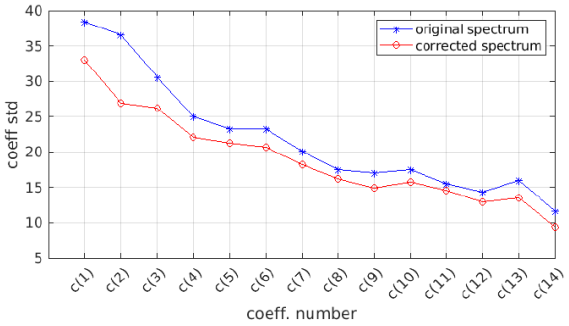


Fig.7. Standard deviations of HFCC cepstral coefficients for the vowel ‘a’

### *D. Global error analysis*

In the subsequent step, for the statistical acoustic models of Polish vowels, the Kulback-Leibler Distance (KLD) was determined. Figure 8 presents the KLD differences between the GMM distributions of the six Polish vowels with and without correction in tabular form. Furthermore, the green colour indicates an increase in the distance (desired) after correction, and the red colour indicates a decrease. In most

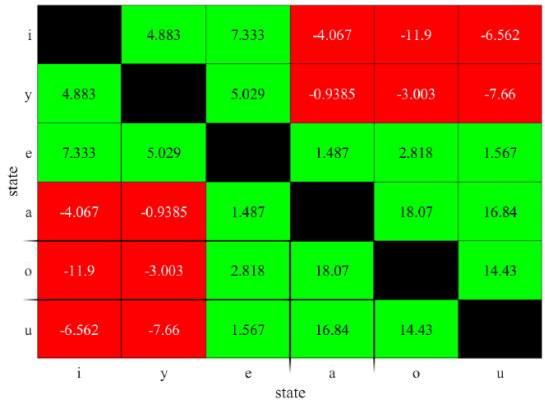


Fig.8. Differences in KLD distances after and before correction between the 6 vowels of Polish speech. The green colour indicates an increase in distance and the red colour a decrease.

of the analyzed vowels, an increase in these distances is observed and the differences are largest for the phonemes a and o. Regrettably, a reduction in the distance between the phonemes ‘i’ and ‘o’ is also evident. This is because the KLD measure uses not only variance but also the mean value vectors of the cepstral coefficient estimators to determine the distance. As mentioned earlier, a reduction in the variance, and consequently greater data concentration, was observed for each coordinate in the observation vectors of all the states analyzed, but for the phonemes ‘i’ and ‘o’ the decrease in variance was small, and a change in the mean value was observed at some coordinates. This change (in the 14-dimensional space) may have reduced the distance between the distributions, in terms of the KLD metric, but is insignificant enough not to have degraded the classification results. Finally, for each analyzed vowel of Polish speech, the FER measure was calculated. A global (for the whole database) analysis of the FER recognition errors at the level of single frames of Polish speech vowels is presented in Fig.9. The results demonstrate a notable reduction

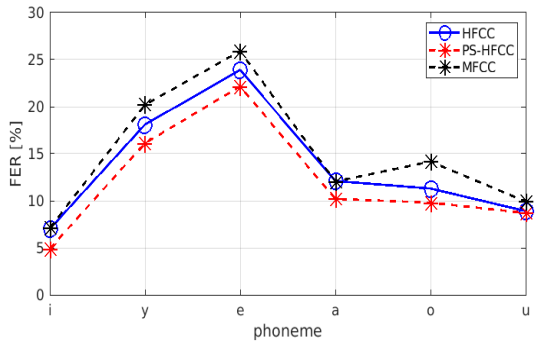


Fig.9. Global FER values for Polish speech vowels. A comparison of classification results obtained using the proposed robust parameterization method, PS-HFCC, with those achieved through the classical HFCC and MFCC approaches.

of recognition errors, indicating the effective reduction of FER errors through the proposed correction. Spectrum correction has been observed to significantly diminish FER values for all of the Polish speech vowels. The classification results obtained using the PS-HFCC parameterization method proposed in this study (expressed in terms of FER) were compared not only with the classical HFCC approach but also with the widely known and commonly used MFCC parameterization method. The error values for individual phonemes, presented in Fig. 9, clearly indicate the effectiveness of employing a variable frame length of the speech signal. While these changes may not be immediately evident, they are nevertheless valuable in speech processing systems with high complexity, where any improvement in parameterization quality is crucial. The proposed approach of pitch synchronization is useful, results in similar classification error reduction, and finds its practical application in other representatives of classical cepstral parametrization methods (i.e. MFCC, GFCC, etc.).

## IV. CONCLUSIONS

The modification of the classical cepstral parameterization methods with pitch-synchronized STFT, as proposed in this paper, has been demonstrated to meet the expected properties. By estimating the fundamental period, $T_0$, and utilizing a variable window length that aligns with the current value of $T_0$, it is possible to effectively eliminate the influence of quasi-periodicity on the amplitude spectrum of voiced speech. Moreover, an observable reduction in the area occupied by the feature vector coordinates can be seen for each vowel, based on the variance of these coordinates. From the other side the ambiguous form of the table with calculated values of the Kullback-Leibler distances between the GMM distributions of Polish speech vowels and improvements in classification errors of individual frames measured by the frame-error-rate measure. In fact the experiments with the proposed method of pitch synchronized cepstral parameterization result in a moderate increase in the efficiency of the vowel classification system, due to the limitations of the GMM model, which generalizes very well, but at the same time is not very sensitive to proposed modifications. It is also worth noting the possibility of modifying the proposed PS-STFT method and combining it with other algorithms known from the literature. One such modification, which utilizes a variable frame length of the signal (synchronization with the fundamental period $T_0$) along with inverse filtering applied to smooth the spectrum and estimate the vocal tract amplitude response, is described in the work [28]. It is still worth keeping in mind that in Automatic Speech Recognition any improvement in the classification stage is valuable. The literature clearly shows that cepstral parameterization methods and the use of the GMM model have been most commonly used in the context of Automatic Speech Recognition for years, but these observations clearly indicate the necessity for further research into the development of new effective classifiers for speech technology systems. It is also important to note that the variability of the components of the feature vector, in addition to the influence of the quasi-periodicity of the glottal excitation, is affected by a number of other factors, including:

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

## REFERENCES

[1] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980. [Online]. Available: https://doi.10.1109/TASSP.1980.1163420

[2] 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

[3] 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

[4] 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

[5] N. Moritz, J. Anemu¨ller, and B. Kollmeier, “An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1926–1937, 2015. [Online]. Available: https://doi.10.1109/TASLP.2015.2456420

[6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, 1993.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Acoustical Society of America Journal*, vol. 87, no. 4, pp. 1738–1752, 1990.

[8] J. Koehler, N. Morgan, H. Hermansky, and H. G. Hirsch, "Integrating rasta-plp into speech recognition," in *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994, pp. I/421–I/424. [Online]. Available: https://doi.10.1109/ICASSP.1994.389266

[9] H. Hermansky and P. Fousek, "Multi-resolution rasta filtering for tandem-based asr," in *Proc. ISCA Interspeech, Lisbon'05*, 2005, p. 361–364.

[10] M. Skowronski and J. Harris, "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

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

[12] B. Atal, "Automatic Speaker Recognition Based on Pitch Contours," *The Journal of the Acoustical Society of America*, vol. 52, no. 6B, p. 1687–1697, 1972.

[13] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (pefac)," in *19th European Signal Processing Conference, EUSIPCO Barcelona'11*, 2011, p. 451–455.

[14] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, vol. 83, no. 1, p. 257–264, 1988. [Online]. Available: https://doi.org/10.1121/1.396427

[15] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*, 2011, p. 1973–1976.

[16] A. Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002. [Online]. Available: https://doi.org/10.1121/1.1458024

[17] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *in Proc. ICASSP2014*, 2014, pp. 659–663. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6853678

[18] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, pp. 1–21, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X19308795

[19] S. Gmyrek, R. Hossa, and R. Makowski, "Amplitude spectrum correction to improve speech signal classification quality," *International Journal of Electronics and Telecommunications*, vol. 70, no. 3, p. 569–574, 2024. [Online]. Available: https://doi.10.24425/ijet.2024.14958

[20] S. Gmyrek, R. Hossa, "Reducing the impact of fundamental frequency on the hfcc parameters of the speech signal," in *2023 Signal Processing Symposium (SPSympo)*, 2023, pp. 49–52.

[21] 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

[22] S. Kullback, "Information theory and statistics," *Dover Publications, New York*, 1968.

[23] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004. [Online]. Available: https://doi.10.1109/JPROC.2003.823141

[24] 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

[25] W. Jassem, *Podstawy fonetyki akustycznej*, ser. Biblioteka mechaniki stosowanej. Pan´stwowe Wydawn. Naukowe, 1973. [Online]. Available: https://books.google.pl/books?id=bCsSGQAACAAJ

[26] C. Bishop, Pattern recognition and machine learning. New York: Springer 2006

[27] R. Makowski, Automatic speech recognition - selected problems [in Polish: Automatyczne rozpoznawanie mowy - wybrane zagadnienia]. Oficyna Wydawnicza Politechniki Wroclawskiej, 2011.

[28] S. Gmyrek, R. Hossa, and R. Makowski, "The Influence of the Amplitude Spectrum Correction in the HFCC Parametrization on the Quality of Speech Signal Frame Classification," *Archives of Acoustics*, vol. 50, no. 1, p. 59-67, 2025. Available: https://doi.10.24425/aoa.2025.153652