

Improving the vowel classification accuracy using varying signal frame length

Stanislaw GMYREK , Robert HOSSA 

Department of Acoustics, Multimedia and Signal Processing, Wrocław University of Science and Technology, Janiszewskiego 11/17, 50-372 Wrocław, Poland.

Corresponding author: Stanislaw GMYREK, email: stanislaw.gmyrek@pwr.edu.pl

Abstract The parts of speech influenced by glottal pulse excitation, the vocal tract, and the speaker's lips shape the voiced components of the speech signal. On the other hand, semantic information in speech is primarily shaped by the vocal tract. However, the irregularity of the glottal excitation's periodicity contributes to a significant dispersion of the parameterization coefficients, introducing fluctuations into the amplitude spectrum. This study proposes a technique to mitigate the impact of this irregularity on the feature vector. It involves using a variable signal frame length synchronized with the fundamental period T_0 and averaging amplitude spectra over a single period to minimize noise effects, smooth out the characteristics, and reduce the estimator variance. By utilizing the derived HFCC parameters, statistical models representing individual Polish vowels were created using mixtures of Gaussian distributions. Additionally, the impact of these correction concepts on the classification accuracy of speech frames containing Polish vowels was examined.

Keywords: automatic speech recognition, robust parameterization, spectrum correction, GMM model.

1. Introduction

In general, the aim of Automatic Speech Recognition (ASR) is to determine the most likely sentence (word sequence) W that transcribes the speech audio A [1,2,3]. The system consists of an Acoustic Model (AM) which takes audio as the input and produces words W as output, and a language model which takes words W as input and generates a sequence of words as output. The AM architecture consists of cascade of 5 elements [2]: (i) Feature Extraction (FE) block which process audio speech A into observations O , (ii) Frame Classification (FC) block with sequence states Q at the output, (iii) Sequence Model (SM) block which produces phonemes L , and (iv) Lexicon Model (LM) of frame classification process with words W at the output as a bridge between the acoustic and language models. In the field of ASR, three distinct classes of solutions can be identified [3]: (i) classical models based on Hidden Markov Models (HMMs), (ii) End-to-End Deep Models, and (iii) Attention based models. In classical HMM-based approaches, two fundamental solutions can be distinguished: the GMM-HMM model [1, 2, 3] and DNN-Deep Models with HMM [4, 5, 6, 7], that utilize discriminative training to minimize words and phoneme error rates. To the second class of ASR belong End-to-End Deep Models with the output of the FE block in the form of waveform or spectrogram, the FC block with RNN (Recurrent Neural Network) [10,11,12] and the SM block based on Connectionist Temporal Classification (CTC) [13]. Finally the third class- Attention based models- plays a crucial role in modeling long sequences. These models directly predict character sequences and simultaneously integrate the FC, SM, and LM blocks, utilizing different forms of acoustic features as input [15,16]. In most of the ASR solutions presented above (including commercial implementations based on patents [8,9,14]), their operation requires a preprocessing stage in the Feature Extraction (FE) block. This stage determines a compact representation for individual segments of the speech signal. The Acoustic Model component of ASR systems must compensate for various undesirable factors affecting speech, including: (i) environmental and technical conditions, (ii) intrapersonal variability (e.g., mood, emotion, health status), (iii) interpersonal variability expressed in differences in age, gender, or speech organ structure, and (iv) contextuality and regional and cultural differences.

In classical GMM-HMM solutions, performance degradation due to these factors can be mitigated by speaker clustering [17,18,19], normalization of parameterization coefficients using cepstral mean and variance normalization (CMVN) [20], and robust parameterization techniques such as RASTA filtering [21]. On the other hand, deep models in both classical and End-to-End ASR systems do not necessarily require

high-level features [3,4]. However, the literature also includes cases where cepstral parameter normalization techniques, such as cepstral mean normalization (CMN) [22], are applied to reduce acoustic channel distortions, as well as global transformation approaches [5]. The present study should be considered in the context of robust parameterization, where the primary objective is to propose low-level feature extraction algorithms that yield parameter sets with low variance and minimal variability under the influence of individual factors, noise, and interference. In particular, our research focused on enhancing the quality of highly distorted STFT spectrograms while simultaneously improving the statistical properties of cepstral parameterization vectors, especially in relation to voiced phonemes. The nature of these phonemes is governed by time-varying periodic excitation (pitch). Unfortunately, this characteristic significantly distorts the amplitude spectrum of such phonemes, which should ideally provide a reliable representation of the instantaneous spectral characteristics of the vocal tract, including easily identifiable formant frequencies.

The mathematical model of the Fant's source-filter type for a discrete-time speech signal $s(n)$ can be expressed as follows [25]:

$$s(n) = v(n) * l(n) * x(n) = h(n) * x(n), \quad (1)$$

where $x(n)$ is the excitation, $v(n)$ the impulse response of the filter modelling the vocal tract, $l(n)$ describes the form of speech emission by a speaker, and $*$ is the discrete convolution operator [3]. The semantic information in speech is primarily influenced by the vocal tract. Conversely, the quasiperiodicity of the glottal excitation adds variability and significant scatter to the resulting coefficients by introducing significant fluctuations in the amplitude spectrum [26]. This is closely related to the signal windowing operation. (see Sect. 2.2).

This paper presents a method to mitigate the impact of glottal excitation through using a varying signal frame length synchronized with the fundamental period T_0 and averaging amplitude spectra over a single period to minimize noise effects, smooth out the characteristics, and reduce the estimator variance (see Sect. 2.3). A key aspect in the proposed method is to compute the fundamental frequency estimator f_0 as precisely and accurately as possible (see Sect. 2.4).

To evaluate the effectiveness of the proposed parameterization methods, statistical models for individual phonemes in Polish speech were developed using a Gaussian Mixture Model (see Sect. 2.5). The goal of the corrections was to narrow the GMM distributions of the amplitude spectrum while increasing the distance between them. According to the detection theory, this approach generally minimizes classification errors. The effectiveness of the corrections was evaluated by comparing Frame Error Rate (FER) measurements before and after applying the correction algorithm (see Sect. 3).

2. Theory

2.1. Short-term feature extraction

Among the many parameterization techniques available, those utilizing time-frequency transforms and cepstral representations are considered some of the most widely used and effective methods [27]. These include Mel Frequency Cepstral Coefficients (MFCC) [28], Human Factor Cepstral Coefficients (HFCC) [29, 30], Basilar-Membrane Frequency-Band Cepstral Coefficients (BFCC) [31], and Gammatone Cepstral Coefficients (GTCC) [32]. In this study, the HFCC representation was selected for its effectiveness in noisy or adverse acoustic conditions, making it valuable for applications in speech and speaker recognition, speech synthesis, and acoustic scene analysis [7,8]. The parameterization produces the feature vector (cepstral coefficients) $c(t,m)$:

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

where $Y_l(t, j)$ represents the logarithm of the signal spectrum in the Equivalent Rectangular Bandwidth (ERB) scale $Y(t, j)$ derived from the amplitude spectrum of speech frame after corrections. Here, t denotes the frame number, j is the frequency band number in the ERB scale, J represents the total number of frequency bands and M is the total number of HFCC coefficients. The HFCC parametrization technique employs a bank of ERB-scaled triangular filters designed to emulate the human auditory system's non-linear frequency perception. This method aggregates spectral energy into frequency bands to align with human hearing characteristics. The energy logarithm in each frequency band is used to mirror the human auditory system's logarithmic perception of loudness. The HFCC approach for extracting speech features is elaborated in details in [30].

2.2. The influence of fundamental frequency f_0 on feature vector

The objects of our interest in this work are the voiced fragments of speech for which the excitation model $x(n)$ takes an impulse form:

$$x(n) = g(n) * p(n) = \sum_{k=0}^{+\infty} g(nT - kT_0) \quad (3)$$

and speech signal $s(n)$:

$$s(n) = \sum_{k=0}^{+\infty} s_p(nT - kT_0), \quad (4)$$

where $g(n)$ is the shape of a single excitation pulse, $p(n) = \sum_{k=0}^{+\infty} \delta(nT - kT_0)$ is a pulse train with a repetition time T_0 (pitch), $s_p(n)$ is the response of the modelling system to a single excitation pulse $\delta(n)$, while T is the sampling interval.

In practical applications, we use a finite-time representation of the speech signal $s(n)$, i.e. $s_w(n)$, which is the result of a windowing operation with the function $w(n)$ of a signal:

$$s_w(n) = s(n) \cdot w(n). \quad (5)$$

As a result of this operation, the spectral representation of the signal analysed in the frame is changed to the form:

$$S_w(\omega) = DTFT\{s_w(n)\} = DTFT\{s(n)\} * DTFT\{w(n)\} = S(\omega) * W(\omega), \quad (6)$$

where $DTFT\{\cdot\}$ is the Discrete Time Fourier Transform (DTFT) operator [23, 24]. In general, the impact of windowing operations in the form of the $W(\omega)$ term can only be compensated for by selecting an appropriate window (e.g. Hamming) and other sophisticated techniques using, for example, amplitude spectrum correction functions determined from inverse filtering methods and estimators of the amplitude of the vocal tract transfer function [33, 34].

For illustration purposes, Fig. 1 presents the amplitude spectra of consecutive frames of the phoneme "a," extracted from longer utterances by the same speaker recorded under identical conditions but differing in fundamental frequencies f_0 . The main difference in these spectral representations lies in the varying locations of the local maxima, which are multiples of the fundamental frequency f_0 . Due to the presence of ripples, the formants are not distinctly visible, although the approximate frequencies of two first formants are 800 Hz and 1.3 kHz. Dotted lines in these figures indicate filterbank with center frequencies corresponding to the mel scale, as used in the HFCC parameterization. The fundamental frequency was calculated using YIN algorithm described in section 2.4.

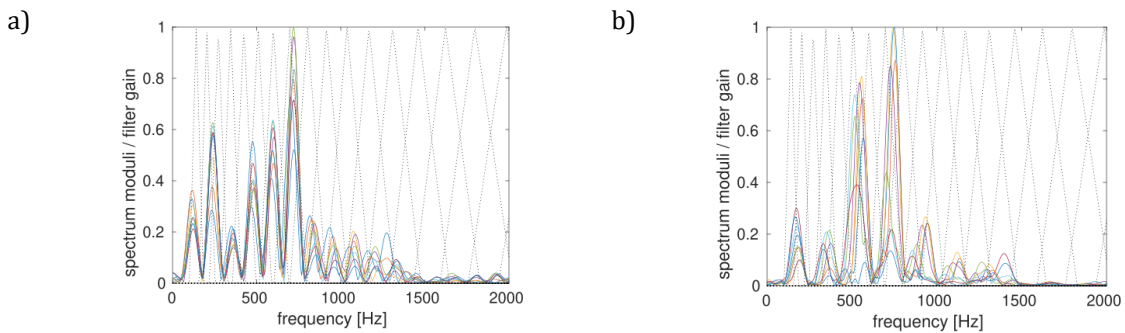


Figure 1. Amplitude spectra of consecutive frames of phoneme 'a' with applied filterbank; the fundamental frequency a) about 130Hz b) about 195Hz. The frequency resolution was 11.72 Hz.

The different positions of the local maxima in the spectrum result in varying energy levels across successive frequency bands, leading to different ERB-scale spectra at different f_0 values. This is illustrated by the ERB-scale spectra plots in Fig. 2, where particularly large differences are observed in band 4.

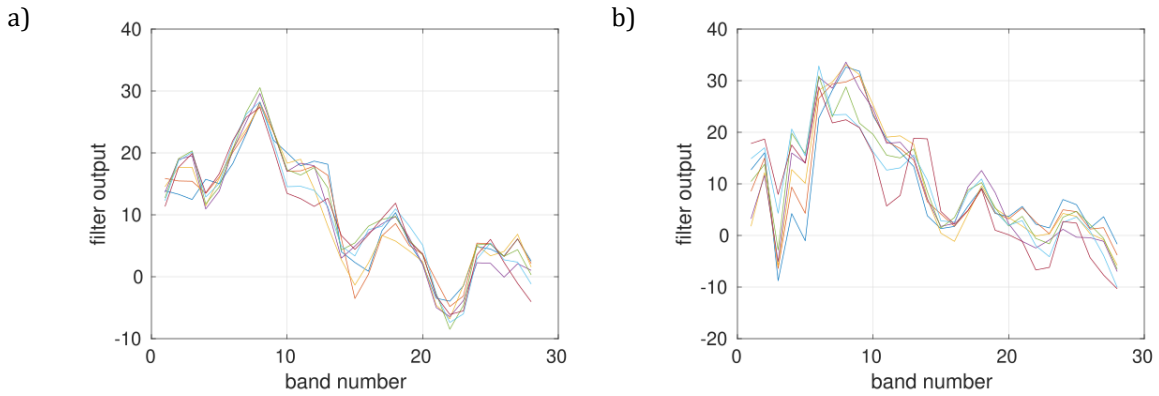


Figure 2. Spectra of consecutive ERB-scale frames of phoneme a; the fundamental frequency is a) about 130Hz, b) about 195Hz.

Consequently, there are significant variations in the cepstral coefficients for the two cases considered, as shown in Fig. 3.

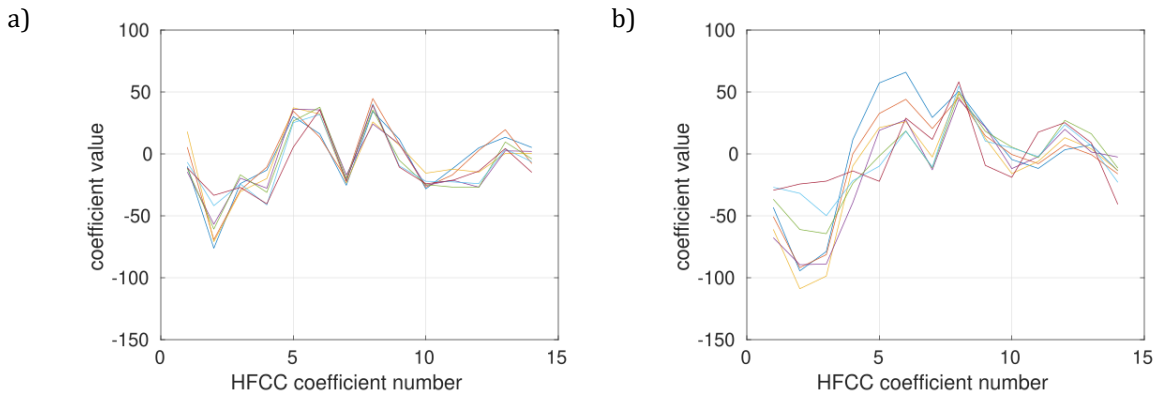


Figure 3. Cepstra of consecutive ERB-scale frames of phoneme a; the fundamental frequency is a) about 130Hz, b) about 195Hz.

2.3. Varying length of the signal analysis window

We limit our further consideration to the special case in which we assume that the length of the analysis window satisfies the condition $T_w = N \cdot T_0$, i.e. is an integral multiple of the period of the fundamental period. In this situation, assuming the local stationarity of T_0 in the frame region, we can write:

$$s_w(n) = \sum_{k=0}^{N-1} s_p(nT - kT_0), \tag{7}$$

which implies the following spectral representation:

$$S_w(\omega) = S(\omega) \cdot \sum_{k=0}^{N-1} e^{-j\omega k T_0} = S(\omega) \frac{\sin(\omega T_0 \cdot \frac{N}{2})}{\sin(\omega T_0 \cdot \frac{1}{2})} \cdot e^{-j\omega \frac{N-1}{2} T_0}, \tag{8}$$

which fully illustrates the problem of significant ripples in the amplitude spectrum of the analysed frame signal $s_w(n)$ even in such a specific case. Fortunately, for $N=1$ the effect of the window on the spectrum of $S_w(\omega)$ completely disappears, but unfortunately implies the need for estimation of the current value of T_0 . In practice, we use the discrete form of the DTFT transform for the speech signal frame, the Discrete Fourier Transform (DFT), and the Fast Fourier Transform (FFT) algorithm is used to calculate it effectively. In addition, for the preservation identical resolution in the discrete spectral representation of successive frames with a varying period length of the fundamental period T_0 , the classical zero-padding technique was used.

2.4. Fundamental frequency estimation

The requirement for full knowledge of the present value of the fundamental period T_0 (or fundamental frequency $f_0 = \frac{1}{T_0}$) enforces the practical use of a simple and efficient method for its estimation. Traditional solutions to the problem of determining the current value of T_0 use the samples from the analysis frame or their power spectrum directly [35]. One of the most commonly used solutions is the YIN algorithm [36], together with its statistically improved version [37]. The version of the algorithm used in the numerical experiments in this paper is based on cumulative mean normalized difference function (step 3 of error reduction in [37]) given in the form:

$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0 \\ d_t(\tau) / \frac{1}{\tau} \sum_{j=1}^{\tau} d_t(j), & \end{cases} \quad (9)$$

where: $d_t(\tau)$ is difference function:

$$d_t(\tau) = \sum_{j=1}^W (x_j - x_{j+\tau})^2 \quad (10)$$

and W is analysis window size. The choice of such a solution ensures the normalization of the $d_t(\tau)$ function and, at the same time, effectively compensates for some of the significant estimation errors especially in the situation of proximity of the fundamental frequency f_0 and the frequency of the first formant of the vocal tract. The use of cumulative mean normalized difference function reduces the errors associated with estimating a doubled value of the fundamental frequency f_0 .

2.5. Correction quality measure

To evaluate the performance of the proposed modifications to HFCC parameterization, a study was conducted on the Polish speech vowels. Creating the above concept necessitated developing acoustic models for these vowels using GMM probability distributions. Evaluation of compensation effectiveness was done using the single frame recognition error measure. At the frame recognition phase, GMM acoustic models comprised a mixture of $K=7$ multivariate normal probability distributions with a diagonal covariance matrix Σ determined through the Expectation-Maximization (EM) algorithm:

$$p_f(o) = \sum_{i=1}^K w_{fi} \mathcal{N}(o, m_{fi}, \Sigma_{fi}), \quad (11)$$

where w_{fi} , m_{fi} denotes the mixture i -th component weights and means for f -th phoneme. The EM algorithm iteratively maximizes the likelihood function of the vectors formed in the process of parameterization of speech signal frames to their statistical model in the form of mixed multivariate normal distributions (GMM). In each step, the algorithm performs operations to average the data vectors and determine their autocovariance matrices for the full GMM model (Maximization step) based on the arrays of conditional probabilities of belonging of these vectors to all components of the estimated GMM mixture previously determined in the Expectation step. The detailed description of the EM algorithm can be found in [38]. Frame Error Rate (FER) is typically used to evaluate the quality of speech recognition at the individual frame level and is defined as

$$FER = \frac{T_{err}}{T} \cdot 100\%, \quad (12)$$

where T is the number of all frames to be recognized and T_{err} is the number of incorrectly recognized frames [26].

3. Experiments and results

The database for the experiments comprises recordings of 36 adult male voices from different Polish cities. Each speaker recorded 150 Polish words, out of which speech segments containing vowels from 43 words were selected for the experiment. The database of recordings consists of single words selected purposely to include a wide range of phonetic contexts, allowing the study of acoustic variation in the speech signal

resulting from a variety of phoneme neighborhoods. The words range from short, monosyllabic forms such as “dwa,” “tak,” and “kot,” to more complex structures such as “czepek,” “żaba” and “zapamiętaj.” Thanks to this diversity, it is possible to capture changes in the parameters of the speech signal, which are determined by both the length of the word and its phonetic structure.

An important element in the construction of the database was the consideration of the statistical properties of the Polish language. Words were selected based on the frequency of occurrence of particular phonemes in Polish speech, in accordance with the results of analyses presented in [41]. Such a selection of material ensures phonetic representativeness and makes it possible to study the signal under conditions similar to natural speech. This is particularly important in the context of speech signal processing and classification, where articulatory and contextual complexity are key challenges.

The recordings had a sampling rate of 12 kHz. The results discussed in this study pertain to recordings with a signal-to-noise ratio of 35 dB. All recordings were manually segmented and labelled based on phonetic units, namely phonemes. The frame length was synchronized with fundamental period T_0 with the 10 ms shift. The number of cepstral coefficients was 14. The speakers were grouped based on cepstral coefficients of the vowels, following a criterion outlined in the paper using the Universal Background Model (UBM) [19].

3.1. Exemplary results

The chapter presents example results of the proposed method of the varying frame length of the speech signal synchronised to the fundamental period T_0 . Figs. 4-5. show the amplitude spectra calculated for several consecutive frames of the speech signal of the phoneme ‘a’ of Polish speech. In particular, Fig. 4. presents the amplitude spectra of the original signal, while Fig. 5. presents the amplitude spectra calculated from a frame of the speech signal, the length of which was selected as a multiple of fundamental period T_0 (according to the methodology presented in Sect. 2.3). The cepstral coefficients were calculated from the two amplitude spectra indicated above. Comparison of Figs. 4-5 shows the evident effectiveness of the proposed method and the removal of the amplitude spectrum ripples caused by the quasi-periodicity of the excitation.

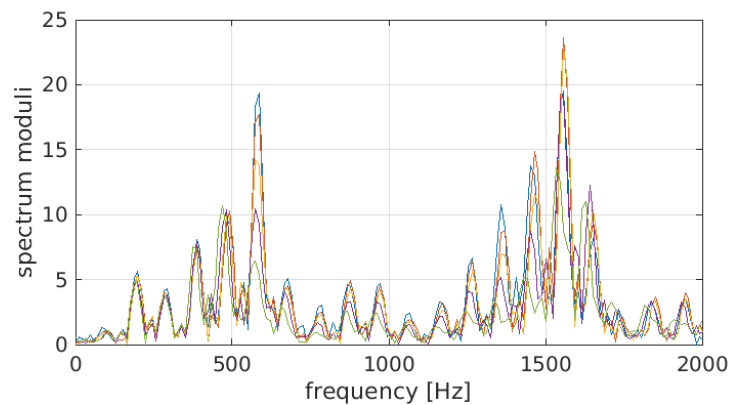


Figure 4. Amplitude spectra of several consecutive frames of phoneme “a” before correction.

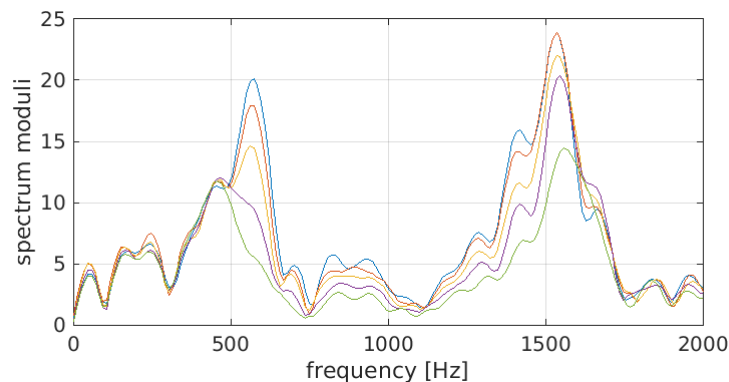


Figure 5. . Example results of the proposed method for several consecutive frames of phoneme “a”.

In turn, Figs 6- 7 show the standard deviations of the values of the individual cepstral coefficients for the two selected states analysed: the vowel ‘e’ (Fig. 6.) and ‘a’(Fig.7) from the whole analysed database.

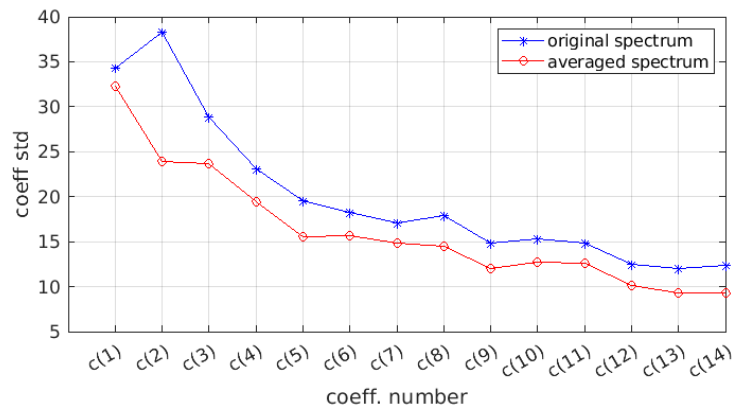


Figure 6. Standard deviations of cepstral coefficient estimators for the vowel ‘e’.

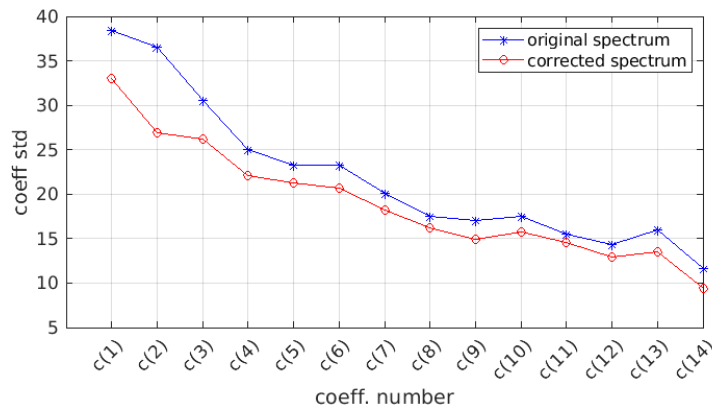


Figure 7. Standard deviations of cepstral coefficient estimators for the vowel ‘a’.

The blue curve indicates the standard deviation of the successive HFCC parameterisation coefficients determined from the amplitude spectrum of the original speech signal, while the red curve indicates the standard deviation of the amplitude spectrum calculated after applying frame length synchronisation to the T_0 . As can easily be seen in the graphs above, the standard deviations shown in the red curves are smaller for each feature vector coefficient.

3.2. Global error analysis

A global (for the whole database) analysis of recognition errors at the level of single frames of the speech signal is presented in Fig. 8. For each analysed vowel of Polish speech, the FER was calculated. To emphasise the correctness and effectiveness of the proposed method, the performance of two classifiers in such conditions was shown: an acoustic model calculated using the GMM model and a decision tree classifier [39]. The blue and black curves show the recognition errors based on the coefficients determined from the original signal, while the red and green curves show the classification errors using the correction method proposed in this paper.

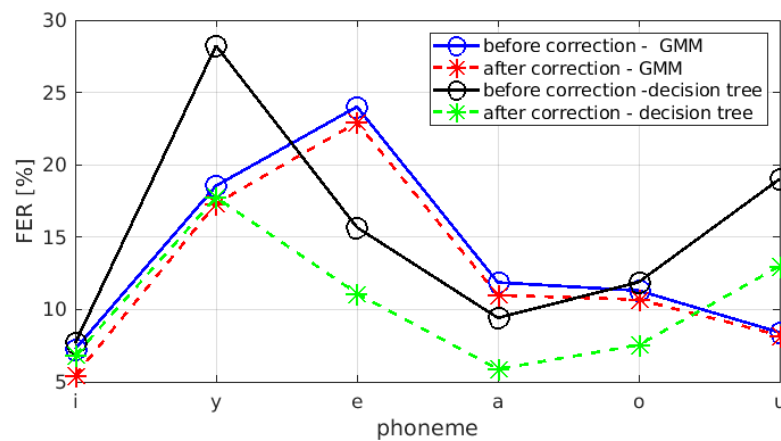


Figure 8. Global FER values for Polish speech vowels.

It is not difficult to see that in both cases improvements were obtained for each condition, nevertheless a more spectacular improvement of classification was obtained using decision trees instead of a few percent of improvement obtained with GMM classifier.

4. Conclusions

The methodology for processing signals representing Polish vowels proposed in this paper, consistent with the instantaneous value of the fundamental period, results in a reduction in the variance of the feature vector's parameters, which leads to an increase in the efficiency of the classification task. The approach discussed in this paper is a continuation of the research proposed in the work [33, 34] and the presented results on real signals show even better recognition quality. The level of classification accuracy was achieved using a significantly smaller number of feature vector parameters than in the traditional approach, which implies a smaller model representation and lower computational complexity of the signal processing algorithms, i.e. reduced resource requirements.

As is well known, a complete ASR system consists of many different subsystems and algorithms. If a few percent increase in recognition accuracy is achieved at each processing stage, with a low computational effort, the total gain in the ASR system will already be significant. It is also worth remembering that the speech signal, due to its randomness and high variability, is also burdened by other factors affecting the feature vector and recognition quality, i.e. inter- and intrapersonal variability, contextuality, technical and environmental factors.

Additional information

The authors declare: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

References

1. L. Deng, M. Lennig, V. Gupta, F. Seitz, P. Mermelstein, P. Kenny; Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition; *IEEE Transactions on Signal Processing*, 1991, 39(7), 1677–1681
2. D. Jurafski, J. Martin; *Speech and language processing: an introduction to natural language processing; Computational linguistics and speech recognition*, Pearson, 2004
3. D. Yu, L. Deng; *Automatic Speech Recognition, A Deep Learning Approach*; Springer-Verlag London, 2015, DOI: 10.1007/978-1-4471-5779-3
4. D. Yu, L. Deng; *Deep Learning Methods and Applications in Foundations and Trends; Signal Processing*, 2022, 7(3-4), 197-387
5. Dahl, G., Yu, D., Deng, L. et al., Context-Dependent Pretrained Deep Neural Networks for Large Vocabulary Speech Recognition; *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(30-42); DOI: 10.1109/TASL.2011.2134090

6. G. Hinton et al., Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups; *IEEE Signal Processing Magazine*, 2012, 29(6), 82-97; DOI: 10.1109/MSP.2012.2205597
7. O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu; Convolutional Neural Networks for Speech Recognition; *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(10), 1533-1545; DOI: 10.1109/TASLP.2014.2339736
8. Patent US8442821B1: Multi-frame prediction for hybrid neural network/hidden Markov models, Google LLC, 2012
9. Patent US8775177B1: Speech Recognition Process, Google LLC 2012
10. H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beau-fays, J. Schalkwyk, Learning Acoustic Frame Labeling for Speech Recognition with Recurrent Neural Networks; *Proc. ICASSP*, 2015, 4280-4284
11. A. Graves, N. Jaitley, Towards End-to-End Speech Recognition with Recurrent Neural Networks; *Proc. of Machine Learning Research*, 2014, 1764-1772.
12. H. Sak, A. Senior, and F. Beaufays, Long short-term memory Recurrent Neural Network architectures for large scale acoustic modeling; *Proc. Interspeech*, 2014, 338-342.
13. A. Graves, S. Fernández, F. Gomez, J. Schmidhuber; Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks; *Proc. Int. Conf. in Learning Representations*, 2006, 369-376
14. Patent US10229672B1, Training acoustic models using connectionist temporal classification, Google LLC, 2017
15. Y. Zhang, W. Chan, N. Jaitly; Very deep convolutional networks for end-to-end speech recognition; 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE Press, 4845-4849; DOI: 10.1109/ICASSP.2017.7953077
16. D. Amodei, et al.; Deep Speech 2: End-to-End Speech Recognition in English and Mandarin; *Proceedings of The 33rd International Conference on Machine Learning; PMLR*, 2016, 48, 173-182
17. T. J. Hazen; A comparison of novel techniques for rapid speaker adaptation; *Speech Communication*, 2000, 31(1); DOI: 10.1016/S0167-6393(99)00059-X
18. R. Kuhn, J-P Junqua, P. Nguyen, N. Niedzielski; Rapid speaker adaptation in eigenvoice space; *IEEE Trans. Speech Audio Processing*, 2000, 8(6), 695-707; DOI: 10.1109/89.876308
19. R. Hossa and R. Makowski; An effective speaker clustering method using ubm and ultra-short training utterances; *Archives of Acoustics*, 2016, 41; DOI: 10.1515/aoa-2016-0011
20. F.H Liu, R.M. Stern, X. Huang, A. Acero; Efficient cepstral normalization for robust speech recognition; *Proceedings of ACL Workshop on Human Language Technologies (ACLHLT)*, 1993, 69-74
21. H. Hermansky, N. Morgan; RASTA processing of the speech; *IEEE Trans. of Speech and Audio*, 1994, 2(4), 587-589; DOI: 10.1109/89.326616
22. O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, *Speech Communication*, 1998, 25, 133-147; DOI: 10.1016/S0167-6393(98)00033-8
23. L. R. Rabiner, R. W. Schafer; *Digital processing of speech signal*; Prentice Hall, 1978
24. A. V. Oppenheim, A. S. Willsky; *Signals & systems (2nd ed.)*. Prentice-Hall, 1996
25. T. F. Quatieri; *Discrete-Time Speech Signal Processing: Principles and Practice*; Prentice Hall, 2001
26. R. Makowski; *Automatyczne rozpoznawanie mowy: wybrane zagadnienia (in Polish)*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2011
27. G. Sharma, K. Umaphathy, S. Krishnan; Trends in audio signal feature extraction methods, *Applied Acoustics*, 2020, 158; DOI: 10.1016/j.specom.2010.04.008
28. S. Davis, P. Mermelstein; Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences; *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980, 28(4), 357-366; DOI: 10.1109/TASSP.1980.1163420
29. M. Skowronski, J. Harris; Improving the filter bank of a classic speech feature extraction algorithm; in *Proceedings of the 2003 International Symposium on Circuits and Systems, ISCAS '03.*, 2003, 4; DOI: 10.1109/ISCAS.2003.1205828
30. M. Skowronski, J. Harris; Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition; *J. Acoust. Soc. Am.*, 2004, 116, 1774-1780 ; DOI: 10.1121/1.1777872
31. T. W. Kuan, A. C. Tsai, P. H. Sung, J. F. Wang, H. S. Kuo; A robust bfcc feature extraction for ASR system; *Artificial Intelligence Research*, 2016, 5; DOI: 10.5430/air.v5n2p14

32. H. Yin, V. Hohmann, C. Nadeu; Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency; *Speech Communication*, 2011, 53(5), 707–715; DOI: 10.1016/j.specom.2010.04.008
33. S. Gmyrek, R. Hossa, R. Makowski; Reducing the impact of fundamental frequency on the HFCC parameters of the speech signal; *2023 Signal Processing Symposium (SPSymposium)*, 2023; DOI: 10.23919/SPSymposium57300.2023.10302705
34. S. Gmyrek, R. Hossa, R. Makowski; Amplitude spectrum correction to improve speech signal classification quality; *International Journal of Electronics and Telecommunications*, 2024, 70(3), 569-574; DOI: 10.24425/ijet.2024.149580
35. W. Hess; *Pitch determination of speech signals*; Springer-Verlag, 1983
36. A. Cheveigne, H. Kawahara; YIN, a fundamental frequency estimator for speech and music, *J. Acoust. Soc. Am.*, 2002, 111(4), 1917-1930; DOI: 10.1121/1.1458024
37. M. Mauch, S. Dixon; PYIN: A fundamental frequency estimator using probabilistic threshold distributions; *Proc. ICASSP2014*, 2014, 659-663; DOI: 10.1109/ICASSP.2014.6853678
38. 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)* ,1977, 39(1), 1–38; URL: <http://www.jstor.org/stable/2984875> (accessed on 2025.06.17)
39. R. Duda, P. Hart, D. Stork; *Pattern Classification*; Willey, 2001
40. C. Basztura; *Rozmawiać z komputerem (in Polish)*; Wyd. Format, 1992

© 2025 by the Authors. Licensee Poznan University of Technology (Poznan, Poland). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).