ARCHIVES OF ACOUSTICS
33, 2, 221–242 (2008)

NORMALIZATION OF SPEAKER INDIVIDUAL CHARACTERISTICS AND COMPENSATION OF LINEAR TRANSMISSION DISTORTIONS IN COMMAND RECOGNITION SYSTEMS

Paweł MRÓWKA[(1)], Ryszard MAKOWSKI[(2)]

[(1)] Neurosoft Sp. z o.o.
Orla 24/1a, 53-143 Wrocław, Poland
e-mail: pawel.mrowka@neurosoft.pl

[(2)] Wrocław University of Technology
Institute of Telecommunications, Teleinformatics and Acoustics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: ryszard.makowski@pwr.wroc.pl

(*received August 01, 2006; accepted February 19, 2008*)

The article presents a novel method of speaker individual characteristics normalization and linear transmission distortion compensation aimed at improving the effectiveness of short isolated utterances recognition. To achieve this goal, spectral transformation banks of a speaker's signal and the division of speakers into classes were applied. The article also discusses the form of spectral transformation, the method of its parameter values optimization, the method of transformation banks definition, the method of speaker classes selection and the way of iterative improvement of recognition results. Moreover, the study puts forward a fast method of speaker classes selection on the basis of the fundamental voice frequency. The efficiency of the proposed solution has been validated by the recognition results obtained by means of four versions of a recognition system using Hidden Markov Models (HMM) and the mel frequency cepstral coefficients (MFCC) parametrization.

Keywords: automatic speech recognition, speaker normalization, transmission distortion compensation.

1. Introduction

The complex challenge facing automatic speech recognition (ASR) has not been adequately coped with yet. The existing systems providing a satisfactory efficiency of recognition suffer from such drawbacks as a small vocabulary and/or the limitation of cooperation to a chosen group of speakers (speaker dependent ASR), or the little robustness to interferences and distortions. It often happens that a system which functions properly for one speaker fails for another one or does not function well in the

case of distortions caused by transmission or recording. ASR systems work on the basis of statistical language models which comprise, among other things, certain acoustic and grammatical qualities. If trained in given acoustic conditions, such a system usually loses some of its efficiency when these conditions change, as they influence the speech signal spectrum. A similar phenomenon can be observed when speakers change, as the time-dependent signal spectrum varies from one speaker to another one. Training speaker and channel independent ASR systems is a solution to this problem and can be done with recordings coming from numerous speakers and at diverse transmission conditions. Still, such a generalizing approach implies a smaller classification capability, which in turn means lower recognition effectiveness.

This is the reason why modern ASR systems contain algorithms of adaptation and/or normalization of the transmission conditions and individual speaker characteristics. Their acoustic models are trained in the way which guarantees the highest classification capability with the use of the aforementioned algorithms. The term "adaptation" is used here with regard to methods which modify parameter values of a model without changing parameters of the analysed signal. Normalization refers to methods altering signal parameters leaving the model parameter values unchanged.

1.1. Review of adaptation and normalization methods

There exist numerous methods of transmission conditions adaptation or normalization. CMN (Cepstral Mean Normalization), CDCN (Codeword Dependent Cepstral Normalization) [1], VTS (Vector Taylor Series) [2] methods rely on signal parameters correction by means of addition of the estimated corrective values. These values are estimated in a way providing the best statistical match of the corrected input parameters to the model trained on the clean signal. Another form of such matching is the cepstral parameters histogram equalization proposed in [3]. The main disadvantage of aforementioned algorithms is the necessity of long (at least a dozen seconds or something like that) input signal fragment analysis, as the input data must be statistically relevant.

Relatively recently missing features methods were devised [4–6]. They cannot be classified in the straight way as the normalization or adaptation algorithms, but their principle is worthy of mention. These methods involve the dividing of the signal into sub-bands and only the sub-bands whose contents have been marked as reliable are taken into consideration in the recognition. The main disadvantage of such methods is the complex and not completely solved task of the reliable sub-band marking.

So far there have been more solutions aimed at the problem of the speaker adaptation than at the channel adaptation, although both issues are related. The VTLN (Vocal Tract Length Normalization) method [7, 8] consists in scaling the frequency axis of the analysed signal spectrum in order to compensate formant frequency shifts caused by differences in the vocal tract length for various speakers.

Algorithms of the MLLR (Maximum Likelihood Linear Regression) type utilize the affine transformation either of the signal parameter space or of the model parameter

space. The transformation can be aimed e.g. at maximization of the classification ability of the model (Conditional MLLR) [9]. A serious disadvantage of the MLLR methods is the number of parameter values to be estimated, that is why some algorithms providing the reduction of this number were proposed, e.g. [10].

A large class of adaptation algorithms is formed by methods of the MAP (Maximum a Posteriori) [11, 12] type which involve the adaptation of model parameter values by using their probability density functions (pdfs) given *a priori* and the available signal parameters of the adapted speaker. And again, these methods need long fragments of the input signal to work well.

Another important group of adaptation algorithms is constituted by methods involving a division of speakers into classes, e.g. the CAT algorithm (Cluster Adaptive Training) [13], in which the model parameter values for a given speaker are determined as the weighted sum of parameters from various classes. On the other hand, the Eigenvoices method [14, 15] performs also the form of soft-clustering, as it analyses the model parameter space by the PCA method. In the adaptation stage, the result model parameters are estimated as a vector in the subspace spanned by PCA output vectors.

The aforementioned adaptation and normalization algorithms are characterized by the following strong regularity: together with the increase in the number of parameters whose values have to be computed, one can notice a simultaneous increase in their efficiency. Unfortunately, this increase means also the necessity of providing fragments of the signal lasting at least a dozen seconds or something like that. The so-called rapid methods, such as Eigenvoices or CAT require signal fragments of at least several seconds. Furthermore, rapid algorithms need training with a training set of at least several tens of speakers.

1.2. The scope of this study

This paper focuses on the compensation of transmission conditions and the normalization of the speaker individual characteristics in recognition of very short and isolated utterances of a duration below one second. Systems of this kind are applied to control devices and it is desirable that they work effectively in various transmission conditions and with various speakers. Due to the very short duration of utterances, the employment of methods known from the literature does not guarantee the achieving of satisfactory results. The reason already mentioned is the too little adaptation data available to apply methods such as CMN, MAP or MLLR. Another reason is the fact, that most of the known methods are based on the iterative adaptation of parameter values. The starting point of the adaptation is of utmost importance here: if the algorithm is initialized improperly, the adaptation becomes inefficient. In the case of short utterances, the small amount of data available increases the probability of inadequate initialization, or, in other words, of wrong initial recognition. Therefore, instead of an iterative method, this study proposes a parallel method based on using spectral transformation banks.

This study is organized as follows: Sec. 2 presents the main principles and the general scheme of the devised algorithm; Sec. 3 describes a form of the proposed spectral

transformation and a method of optimization of the corresponding parameters; Sec. 4 discusses a method of dividing speakers into classes and a method of determining pdfs of mel frequency cepstral coefficients (MFCC) within these classes; Sec. 5 presents a method of the spectral transformation banks construction; Sec. 6 focuses on an algorithm of fast selection of speaker classes on the basis of analysis of the fundamental voice frequency; Sec. 7 presents the variants of the ASR system used in the research and the manner of implementing the recognition result improvement in them; Sec. 8 discusses the results obtained.

2. Principles and general scheme of the recognition improvement method

The algorithm devised is designed to cooperate with the ASR system with a small vocabulary. The normalization will concern such individual characteristics as the timbre of voice and shifts in formant frequencies related to differences in the vocal tract length. The compensation of transmission distortions will encompass constant in time linear distortions of smooth frequency response magnitude, such as distortions coming from microphones and other elements of the electroacoustic path or from the characteristic of the speaker's mouth radiation related to his or her position in relation to the microphone. It was assumed that there were no distortions that would totally remove information from signal sub-bands or no strong reverberation resulting in a deep comb filtering [16] and distorting the time structure of the signal by causing delays comparable to the duration of the analysis frame. However, in most residential and office spaces or inside cars this kind of reverberation does not appear. It was also assumed that the noise level is so low that its influence can be neglected. The general outline of the proposed method is shown in Fig. 1.

Each 20 ms frame of the discrete signal, with the sampling frequency $f_s = 16$ kHz, after windowing and determining of the discrete amplitude spectrum $\mathbf{s}_0$ of length $L = 256$ (the bin for $f_s/2$ was not taken into consideration) is subjected to transformation in spectral transformation (ST) banks. The transformation involves scaling of the frequency axis by $g(f)$ functions banks and linear filtering by LF banks. Each of the M speaker classes can have a different ST bank. Then the MFCC parameterization is carried out together with the observation probabilities $P(\mathbf{o}_t)$[(1)] computing with MFCC pdfs separate for each class (see Sec. 4) and approximated with the sum of 5 Gaussians (GMM). A phoneme was the basic language unit being modeled. We used 35 phonemes including 6 vowels. The next stage was utterance recognition. We chose a $\{m, k_g, k_{LF}\}$ combination (m – speaker class, k_g – $g(f)$ function, k_{LF} – LF) that guarantees the highest value of the recognition score defined in the system (see Sec. 7). This recognition score has a function similar to the confidence measure. To improve recognition, we also deployed an iterative algorithm increasing the variety of $g(f)$ functions and LF used within the version $\{m, k_g, k_{LF}\}$ chosen earlier. Moreover, we used the estima-

[(1)] $P(\mathbf{o}_t)$ is an abbreviated marking of observation probabilities of HMM states for vectors $\mathbf{o}_t$ of MFCC parameters, for t = 1...T, where T stands for utterance duration in frames.

tion of fundamental frequency F_0 to reduce the computational cost by the preliminary assignment of the speaker to a proper class on the basis of the F_0 value.

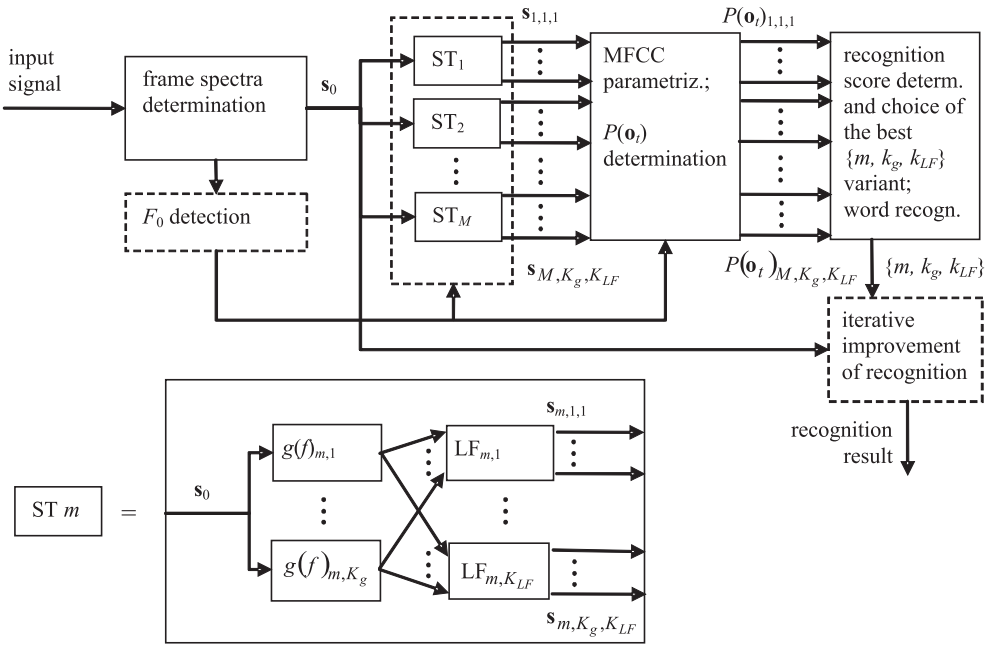

Fig. 1. The outline of transmission distortions compensation and speaker individual characteristics normalization, ST – spectral transformation, $g(f)$ – frequency axis scaling function, LF – linear filter, F_0 – fundamental voice frequency.

Figure 2 shows the general outline of the ST banks construction and speaker classes determination method. Before the division of speakers into classes takes place, one can optionally take into account spectral transformations for speakers, determined for each speaker from the training set in relation to any other member of this set. After designating the speaker classes there is also a possibility of carrying out transformations aimed at increasing the classification capability of MFCC pdfs determined within the classes in the following stage. The parameter values of these transformations serve also as one of the initial parameters in the process of determining final transformations of speaker spectra within the classes. The sets of parameter values of final transformations serve as the basis of the ST banks construction.

We used the "CORPORA" database of the Polish language [17], which includes 45 recording sets done by 37 speakers (25 males, 12 females) of age from 9 to 70 years. Each recording consists of 365 utterances (200 first names, 33 alphabet letters, 10 digits, 8 control commands and 114 short sentences). The recordings were made in a plain office environment. A condenser microphone was used and the digitalization with parameters $f_s = 16$ kHz, 12 bits/sample were done. The base is also segmented and labeled (phoneme is the unit).

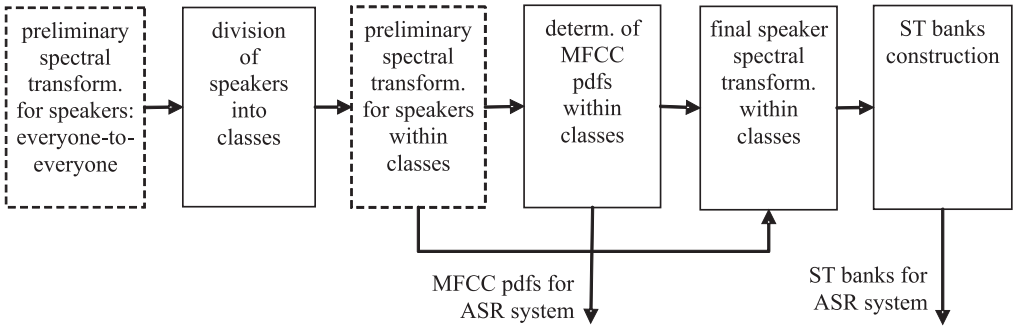

Fig. 2. The outline of the ST banks construction and speaker classes determination method.

For our research, we divided the database into the training set consisting of 25 sets of recordings (20 speakers) and the test set of 20 recordings (17 speakers).

3. Spectral transformation

The scaling of the frequency axis and linear filtering are conducted on the signal amplitude spectrum. Considerable advantages of this method are the small number of parameters whose values have to be determined, the physical meaning of its parameters, its universal quality and the relative independence from the training set. The aforementioned physical meaning of the parameters matters in such analyses as the interpolation of parameters carried out in the proposed iterative improvement of recognition (see Sec. 7). The relative independence from the data of the training set is desirable when this set is small. In the case of sets employed in the research the postulated method yields good results for speakers outside the training set, whereas the Eigenvoices method proves ineffective for such speakers as shown in [18].

3.1. The transformations applied in the method

It was assumed that the scaling function $f_b = g(f)$ is piecewise linear (Fig. 3), where f_b and f stand for the frequency before and after the transformation, respectively. The parameters $f_1, \ldots, f_N$ and $f_{b\max} = f_{\max}$ are constant, whereas the values of the $f_{b\,1}, \ldots, f_{b\,N}$ parameters are established through optimization. They are also constrained in such a way as to guarantee the monotonicity of $g(f)$. In the implementation it was assumed that $f_1 = 1.4$ kHz, $f_2 = 2.3$ kHz, $f_3 = 4.1$ kHz and $f_{\max} = 8.0$ kHz.

The linear filtering was performed by the minimum-phase FIR filter. The location of the transfer function zeros was limited so that they did not lie too close to the unit circle, which in turn prevents too high local attenuation. In the implementation it was assumed that the number of zeros is 4 and that their maximal radius is 0.8. The radii and angles delineating the position of zeros are determined by optimization.

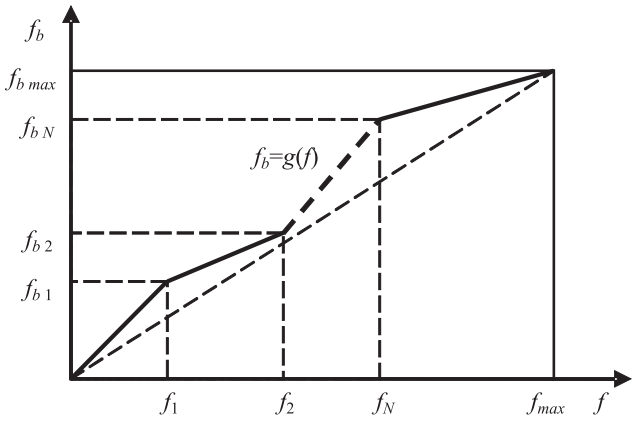

Fig. 3. Scaling function $g(f)$.

3.2. Optimization of transformation parameter values

Transformation parameter values for a given speaker are set optimally, while the objective of optimization is to maximize the recognition result of isolated signal frames of this speaker. This in turn ensures the universality and independence of the obtained transformations from the current contents of the system vocabulary. The recognition of isolated frames is performed on the basis of the given joint MFCC pdfs henceforth referred to as reference pdfs. The determination of the transformation parameter values is carried out in several points of the algorithm. Depending on these particular points, the reference pdfs may be pdfs for speaker classes or individual pdfs for a single speaker. In the case of switching from marginal pdfs to joint pdfs, it was assumed that they are statistically independent. Segmentation and labeling available in the database were used. The recognition result for vowels was maximized with the simultaneous limitation preventing a parallel decrease in recognition for other phonemes. The reason for such a procedure was that the adequate recognition of vowels is crucial in recognizing short utterances from a small vocabulary.

The starting point for defining the optimization objective function is the expected value of the error in recognizing isolated frames for a given phoneme r – relation (1). The error is understood here as the difference of the log-likelihood obtained for phoneme r and the maximum log-likelihood obtained for other phonemes, weighted the function w:

$$c_r = \int\limits_{\Re^D} w \left(\ln\left(p_{2,r}\left(\mathbf{o}\right)\right) - \max_{\substack{j=1\ldots R \\ j\neq r}} \left\{ \ln\left(p_{2,j}\left(\mathbf{o}\right)\right) \right\} \right) p_{1,r}\left(\mathbf{o}\right) \mathrm{d}\mathbf{o}, \tag{1}$$

where D denotes the dimension of the vector $\mathbf{o}$ (the number of MFCC), $p_{1,r}$ is the joint MFCC pdf for the phoneme r of the speaker whose spectrum is being transformed (this pdf depends on the transformation parameter values and changes in the process of optimization), $p_{2,r}$ is a constant joint reference pdf for the phoneme r. The following

versions of the function w were proposed: $w_1(x) = x$ and $w_2(x) = u(x)$, where $u(x)$ denotes a unit step function.

The value of function (1) may be estimated by the following relation:

$$\widetilde{c}_r = \frac{1}{N}\sum_{n=1}^{N} w\left(\ln\left(p_{2,r}\left(\mathbf{o}_{n,r}\right)\right) - \max_{\substack{j=1\ldots R\\ j\neq r}}\left\{\ln\left(p_{2,j}\left(\mathbf{o}_{n,r}\right)\right)\right\}\right), \tag{2}$$

where N denotes the number of the vectors $\mathbf{o}_{n,r}$, which are MFCC vectors of a given speaker for the phoneme r obtained from signal frame spectra after spectral transformation with certain values of its parameters changed in the process of optimization. Hence, the vectors $\mathbf{o}_{n,r}$ change during optimization. The estimator (2) is unbiased and consistent.

Let C_{vow} denote the average value of $\widetilde{c}_r$ for vowels and C_{con} stand for the average value for consonants. C_{conb} takes on the value C_{con} computed at the start of optimization: in version 1 – for neutral parameter values of spectral transformation (no transformation); in version 2 – for parameter values of a certain preliminary transformation defined in the course of selecting speaker classes. The following objective functions were proposed:

$$C_1 = \begin{cases} C_{\text{vov}}, & C_{\text{con}} \geq C_{\text{conb}}, \\ C_{\text{vow}} - \alpha_1 \cdot \left(C_{\text{conb}} - C_{\text{con}}\right), & C_{\text{con}} < C_{\text{conb}}, \end{cases} \tag{3}$$

for the function w in (2) equal w_1 and

$$C_2 = \begin{cases} C_{\text{vow}}, & C_{\text{con}} \geq C_{\text{conb}}, \\ C_{\text{vow}} \cdot \alpha_2^{C_{\text{conb}} - C_{\text{con}}}, & C_{\text{con}} < C_{\text{conb}}, \end{cases} \tag{4}$$

for the function w in (2) equal w_2. In the implementation we set penalty coefficients $\alpha_1 = 2$ and $\alpha_2 = 0.87$.

From (2), (3), and (4) it follows that in the case of the function C_2 we calculate the recognition result of isolated frames, whereas in the case of function C_1, we also take into account the dynamics of recognition errors understood herein as the difference of log-likelihoods.

The estimator (2) is applied to objective functions C_1 and C_2 and their values depend on the chosen set of frames used to determine the vectors $\mathbf{o}_{n,r}$. Hence, the value of the objective function is influenced by the estimation error. We made attempts at using a stochastic optimization algorithm [19] where sets of frames are drawn in each iteration. However, this method did not yield satisfactory results and finally just one set of frames was chosen at the start of optimization. The set was rich enough to make the estimation error much smaller than the obtained recognition result improvement and to come up with an appropriate generalization, i.e., the obtained improvement was also maintained for other selected sets of frames.

Owing to the character of the objective functions used – they can be discontinuous and have many local maxima – a two-stage optimization algorithm was proposed. In the first stage an evolutionary algorithm [20] is applied to find a region in which the

global maximum is supposed to be located. In the second stage the Nelder-Mead simplex method is employed to find the exact location of the maximum. A similar proposition of a hybrid algorithm can be found e.g. in [21].

4. Division of speakers into classes

The recognition based on using one MFCC pdf for all speakers are not efficient enough due to the considerable variability of the speakers characteristics. On the other hand, the use of numerous pdfs coming from individual speakers does not enable an effective normalization of other speakers' individual characteristics by means of the proposed spectral transformation method. Thus it is justified to resort to an transitional solution and to use pdfs for speaker classes. The suggested methods of the speaker classes selection and MFCC pdfs determination are described below. Owing to the small size of the training set, the speakers were divided into two classes.

Let us assume that there are N $\{\nu_1, \ldots, \nu_N\}$ speakers in the training set and that we know the values of the similarity measure between speakers i and j denoted as $d_{i,j}$. We also know the number of classes M ($M < N$) to which all speakers will be assigned. In order to designate speaker classes, in which there is a specific central speaker ν_m, for each combination q: $\{\nu_1^q, \ldots, \nu_m^q, \ldots, \nu_M^q\}$, where ν_m^q are those selected from among N speakers, we compute

$$d_{\text{sum}}^q = \sum_{n=1}^{N} \left\{ d_{j,n} : \; j = \underset{m=1\ldots M}{\arg\max} \left\{ d_{\nu_m^q, n} \right\} \right\} \tag{5}$$

and then we find the number of the optimal combination q_{opt}, for which d_{sum}^q took on the highest value. The set of speakers $(\nu_1^{q\,\text{opt}}, \ldots, \nu_M^{q\,\text{opt}})$ is taken as the class centers. The remaining speakers are assigned to classes so that a given speaker belongs to a class in which his or her similarity to the central speaker has the highest value.

4.1. Version 1 of the speaker classes determination method

The similarity measure among the speakers has the following representation:

$$d_{i,j} = \frac{1}{R} \sum_{r=1}^{R} \prod_{d=1}^{D} \frac{\int\limits_{\Re} p_{d,r,i}(x) \cdot p_{d,r,j}(x)\, \mathrm{d}x}{\sqrt{\int\limits_{\Re} p_{d,r,i}^2(x)\, \mathrm{d}x \cdot \int\limits_{\Re} p_{d,r,j}^2(x)\, \mathrm{d}x}}, \tag{6}$$

where $p_{d,r,i}$ denotes pdf of the d-th MFCC coefficient for phoneme r for the speaker i.

4.2. Version 2 of the speaker classes determination method

The speaker classes were designated on the basis of the measure (6). Next, the preliminary optimizations of spectral transformation parameter values for all speakers of

a given class were carried out, with the reference pdfs equal to those of the central speaker in the class. The objective function C_2 and version 1 of determining value C_{conb} were used in these optimizations. The MFCC pdfs for classes were determined with the obtained preliminary transformations taken into consideration.

4.3. Version 3 of the speaker classes determination method

The similarity measure (6) does not take into consideration the classification capability of pdfs being compared, as it is only a similarity measure of the same phonemes of two speakers. Hence a distance between speakers was introduced[(2)]. It takes into account not only the similarity between the same phonemes but also the distance from other phonemes:

$$d_{i,j} = \sum_{r=1}^{R} \frac{b_{i,j,r}}{\sum\limits_{\substack{k=1 \\ k \neq r}}^{R} b_{i,j,k}}, \qquad b_{i,j,r} = -\ln\left(\int\limits_{\Re^D} \sqrt{p_{r,i}\left(\mathbf{o}\right) \cdot p_{r,j}\left(\mathbf{o}\right)} \mathrm{d}\mathbf{o} \right), \tag{7}$$

where $p_{r,i}$ denotes the joint D-dimensional MFCC pdf for phoneme r for the speaker i, and $b_{i,j,r}$ is the Bhattacharyya distance between pdfs $p_{r,i}$ and $p_{r,j}$. The Bhattacharyya distance is often employed to compare pdfs [22, 23]. Distance (7) decreases along with the increase in similarity among speakers. Hence in (5) the search for the maximum should be replaced with the search for the minimum. One of the properties of the distance (7) is the normalization of the influence of different phonemes r.

It was suggested that the division into classes should be done after conducting preliminary speaker spectral transformations of the everyone-to-everyone type. Hence the values $d_{i,j}$ are determined with these transformations taken into consideration. In the transformation parameter values optimization the distance (7) was applied in the objective function.

After the designation of speaker classes and before the determination of the final MFCC pdfs, a one more optimization is conducted. The speaker spectra in a given class are transformed so as to minimize the objective function which is the sum of the values of distance (7) for each speaker in the class, excluding the central speaker, to the average pdf in the class determined on the basis of the data from speakers after the spectral transformation, and constantly updated in the course of optimization. The final MFCC pdfs for the classes were determined with regard to the transformations obtained.

5. Spectral transformation banks

The LF banks and $g(f)$ banks are determined independently and separately for each speaker class. The input data for bank construction algorithms are the sets of spectral transformation parameter values determined for each speaker from the training set.

[(2)] The term "distance" was used here despite the fact that it does not follow the triangle inequality.

5.1. Determination of bank elements

The method of hierarchical clustering, with Ward's distance between the clusters, was applied to determine the elements of the ST bank. Ward's distance allows for linking clusters at each step of the algorithm operation in such a way that the total error, related to the approximation of the elements undergoing clustering by cluster centroids, grows minimally. The scheme of the algorithm is as follows:

Let **x** denote a vector containing transformation parameters that undergo clustering. The same algorithm is used both for $g(f)$ (then **x** consists of parameters $\{f_{bi}\colon i = 1, 2, 3\}$), and for LF (then $\mathbf{x} = \mathbf{h}$, where **h** stands for filter frequency response magnitude). Let K denote the target number of clusters and N stand for the number of vectors **x** undergoing clustering.

1. Initialization. From $N > K$ of the input vectors **x** create N one-element clusters.

2. Compute Ward's distance for each cluster pair i, j:

$$\mu_{i,j} = \sum_{n=1}^{N_i+N_j} \left\|\mathbf{x}_n^{i+j} - \mathbf{x}_c^{i+j}\right\|_L^2 - \left(\sum_{n=1}^{N_i} \left\|\mathbf{x}_n^{i} - \mathbf{x}_c^{i}\right\|_L^2 + \sum_{n=1}^{N_j} \left\|\mathbf{x}_n^{j} - \mathbf{x}_c^{j}\right\|_L^2\right), \quad (8)$$

where **x** with the index n denotes the n-th element, and with the index c – stands for the centroid of clusters i or j or of the cluster resulting from the joining of clusters i and j. The centroid is determined as the arithmetic mean of elements in a given cluster. N_i stands for the number of elements in cluster i; analogously, N_j. $||\bullet||_L^2$ denotes the squared modified Euclidean norm described below.

3. Link clusters i and j of the least $\mu_{i,j}$ obtained and decrement N by 1.

4. If $N > K$, return to step **2**. Otherwise, terminate the algorithm and take the centroids of the obtained clusters as bank elements. In the case of filter clustering, approximate the obtained frequency response magnitude by the filter parameters of the given order.

The squared modified Euclidean norm used in (8) was defined as:

$$\|\mathbf{x}\|_L^2 = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{L} \cdot \mathbf{L}^{\mathrm{T}} \cdot \mathbf{x} = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{D} \cdot \mathbf{x} = (\mathbf{V} \cdot \mathbf{x})^{\mathrm{T}} \cdot (\mathbf{V} \cdot \mathbf{x}), \quad (9)$$

where **V** is a full rank matrix transforming the parameter space. Hence **D** is a positive-definite matrix which can be factorized by means of the Cholesky factorization into two lower-triangular matrices **L**. For the function $g(f)$ matrix **L** has the size of 3 and the values of all non-zero elements are determined in the optimization presented below. In the case of filters matrix **L** has the size of 256. To decrease the number of optimized parameters, one determines only the values of the parameters located on the main diagonal (building vector **l**). Further, the number of optimized values is reduced by approximation of vector **l** in the cosine basis.

In the optimization, the objective function was minimized:

$$l(\mathbf{L}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \left(\|\mathbf{x}_n - \mathbf{x}_m\|_L^2 - (C_{n,n} - C_{n,m})\right)^2, \quad (10)$$

where N denotes simultaneously the number of input vectors **x** for the bank construction algorithm and the number of speakers in a given class, as every **x** is a vector of spectral transformation for a given speaker. Matrix **L** was optimized separately for each speaker class and each version of the determination of spectral transformation parameters. $C_{n,m}$ denotes the value of the objective function used in determining spectral trans-

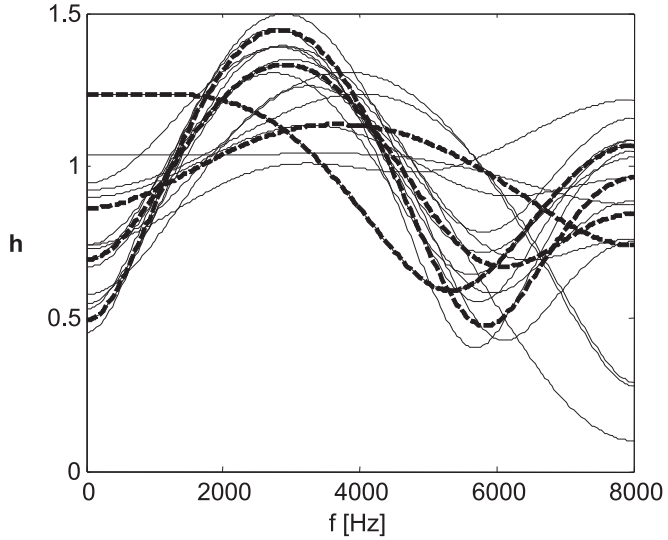

Fig. 4. Examples of LF frequency response magnitudes for speakers (continuous lines) and elements of the LF bank determined on this basis (dashed lines).

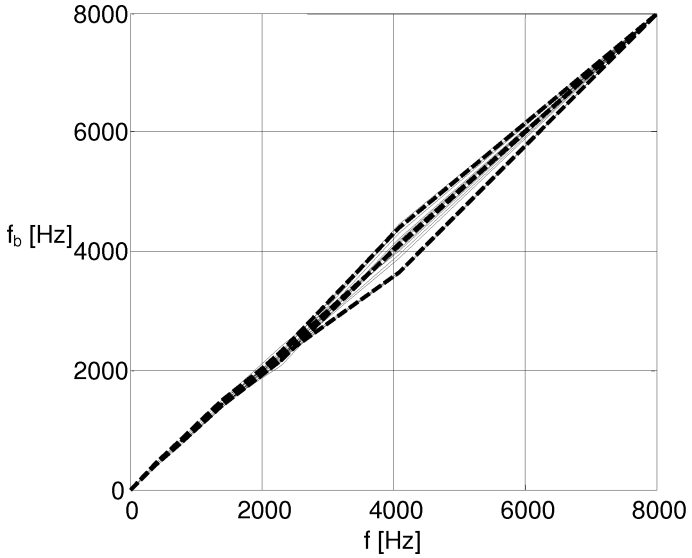

Fig. 5. Examples of functions $g(f)$ for the speakers (continuous lines) and elements of function $g(f)$ banks determined on this basis (dashed lines).

formation parameters obtained for speaker n, whose spectrum was transformed with the parameters $\mathbf{x}_m$. The reference pdfs used in determining the value $C_{n,m}$ were the pdfs for a given speaker class. Depending on the version, functions C_1 or C_2 were applied. Owing to the form of function (10), the Levenberg–Marquardt algorithm was used [24]. The goal of the optimization presented here is to increase the correlation coefficient between differences in the recognition result for isolated frames and the squared norm of the difference of spectral transformation parameter vectors, as the intended use of ST banks is to enhance recognition. The following results were obtained: the correlation coefficient increased from 0.64 to 0.80 for parameters describing LF and from 0.61 do 0.84 for parameters defining $g(f)$.

In the implementation the target number of bank elements was $K = K_g = K_{LF} = 4$. Figures 4 and 5 show examples of LF frequency response magnitudes and $g(f)$ determined for speakers, together with elements of ST banks constructed on their basis.

5.2. Transmission distortions taken into account

The LF banks determined in the manner presented above did not guarantee a satisfactory compensation of channel distortions. Therefore, a modification consisting in cascade joining of compensations for transmission differences and differences between speakers was introduced into the algorithm. Five frequency response magnitudes of popular microphones were approximated by means of the AR model of the 7th order, with the application of the Yule–Walker method [25]. They were marked as $\mathbf{q}_1$ to $\mathbf{q}_5$. Three of them were used in the training procedure, whereas all 5 were used in the tests. Additionally, two linear characteristics of the slope +6 dB/8 kHz and -6 dB/8 kHz were created and marked as $\mathbf{q}_6$ and $\mathbf{q}_7$, respectively. The model changes in the characteristics

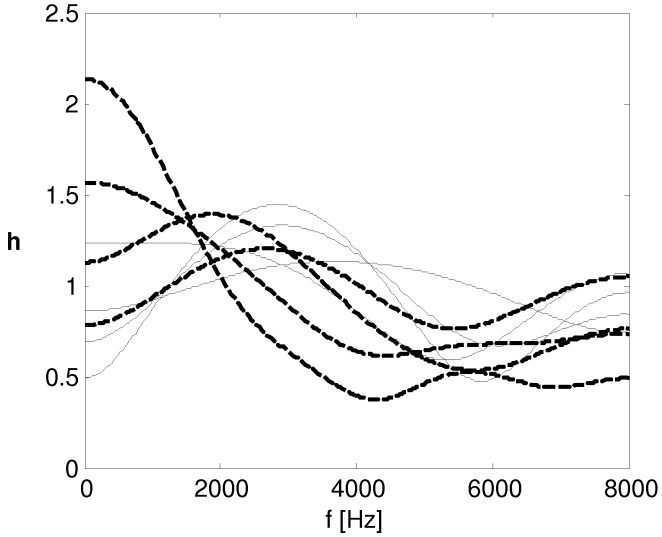

Fig. 6. Examples of elements of the LF bank determined without the modification taking into consideration transmission distortions (continuous lines) and with such a modification (dashed lines).

of the speaker mouth radiation and in changes related to directional frequency response magnitudes of the microphones. It was assumed later that $\mathbf{q}_0$ denotes the frequency response magnitude of the unit transfer function and that $\mathbf{q}_i^{-1}$ is the inverse of $\mathbf{q}_i$.

The modification of the filter banks determination method consisted in completing the set of N frequency response magnitudes $\mathbf{h}_n$, constituting the input to the bank construction algorithm, with those modified by modeled linear distortions. The completed set has the following representation:

$$\left\{ \mathbf{h}_n \circ \mathbf{q}_i^{-1} \circ \mathbf{q}_j^{-1} : \ n = 1, ..., N, \ i = 0, 1, 2, 3, \ \ j = 0, 6, 7 \right\}. \tag{11}$$

The operation '$\circ$' in (11) is understood as the element-wise vector multiplication. Figure 6 shows elements of LF banks obtained from the set of input characteristics before and after the modification of this set.

5.3. Additional bank elements determination

Even when one employs hierarchic clustering, it is impossible to increase the number of filters or functions $g(f)$ available on its lower levels, since it is limited by the number of speakers in the training set. A more substantial variety of filters and $g(f)$ functions was achieved by means of the iterative method. For each cluster resulting from the algorithm in Subsec. 5.1, two additional elements $\mathbf{x}_{d1}$ and $\mathbf{x}_{d2}$, were determined by means of the method given below. The algorithm of determination of additional elements employs direct search as it should correspond to the algorithm applied at the stage of the proper recognition by the iterative method. The direct search was used there with regard to the properties of the objective function (see Sec. 7).

1. Initialization. Set $i = 0$.

If a given cluster contains more than 1 element, set:

$$\mathbf{x}_{d1} = \mathbf{x}_c + \lambda^{1/2}\mathbf{v},$$
$$\mathbf{x}_{d2} = \mathbf{x}_c - \lambda^{1/2}\mathbf{v},$$

where λ and $\mathbf{v}$ denote, respectively, the greatest eigenvalue and the corresponding eigenvector of the covariance matrix of elements $\mathbf{x}_n$ of a given cluster, and $\mathbf{x}_c$ stands for the centroid of this cluster. Set matrix $\mathbf{B}_0$ to the matrix containing elements $\mathbf{x}_c$, $\mathbf{x}_{d1}$ and $\mathbf{x}_{d2}$ (row vectors) in its subsequent rows. Proceed to step **2**.

If a given cluster contains one element, set:

a) for filters:

$$x_{d1,m} = (0.8 + 0.4 \cdot m/L) \cdot x_{c,m},$$
$$x_{d2,m} = (1.2 - 0.4 \cdot m/L) \cdot x_{c,m},$$

for $m = 0 \ldots L - 1$, where $x_{\cdot,m}$ denotes the $m-$th element of vector $\mathbf{x}$.

b) for $g(f)$ functions:

$$\mathbf{x}_{d1} = 1.1 \cdot \mathbf{x}_c,$$
$$\mathbf{x}_{d2} = 0.9 \cdot \mathbf{x}_c.$$

Terminate the algorithm.

2. Increment i by 1. Set $j = 1$. Repeat steps **3**–**6** for each element $\mathbf{x}_n$ of the cluster.
3. Set matrix $\mathbf{A}_j$ to the identity matrix of size 3.
4. Compute the distance (9) of the cluster element $\mathbf{x}_n$ to $3+2(j-1)$ elements which are rows of the matrix $\mathbf{E} = \mathbf{A}_j \cdot \mathbf{B}_{i-1}$. Mark the three rows of matrix $\mathbf{E}$, for which minimal distances were obtained, according to the increasing distance order as k_1, k_2 and k_3. Construct matrix $\mathbf{A}_{j+1}$ by adding two rows to matrix $\mathbf{A}_j$. They have to equal the means of rows of matrix $\mathbf{A}_j$ indexed as k_1 and k_2 and as k_1 and k_3. This corresponds to creating two new points of the direct search.
5. If $j < J$, increment j by 1 and return to step **4**.
6. Obtain the number of a row of matrix $\mathbf{E} = \mathbf{A}_{j+1} \cdot \mathbf{B}_{i-1}$ satisfying the following condition: the distance (9) from the element in this row to the element $\mathbf{x}_n$ of the cluster is the shortest one. Save the row of matrix $\mathbf{A}_{j+1}$ with this number and mark it as $\mathbf{a}_n$.
7. Determine a new matrix $\mathbf{B}_i$ by solving the following well-determined or over-determined set of equations by means of the least squares method.

$$\mathbf{A}^{(r)}\mathbf{B}_i^{(r)} = \mathbf{X} - \mathbf{a}^{(r)}\mathbf{x}_c, \quad \mathbf{B}_i = \left[\mathbf{x}_c^{\mathrm{T}}\mathbf{B}_i^{(r)\mathrm{T}}\right]^{\mathrm{T}}, \tag{12}$$

where matrix $\mathbf{A}^{(r)}$ consists of subsequent rows $\mathbf{a}_n$, in which the first element was omitted. The first elements of vectors $\mathbf{a}_n$ build a column vector $\mathbf{a}^{(r)}$. Matrix $\mathbf{X}$ contains cluster elements $\mathbf{x}_n$ in its subsequent rows.

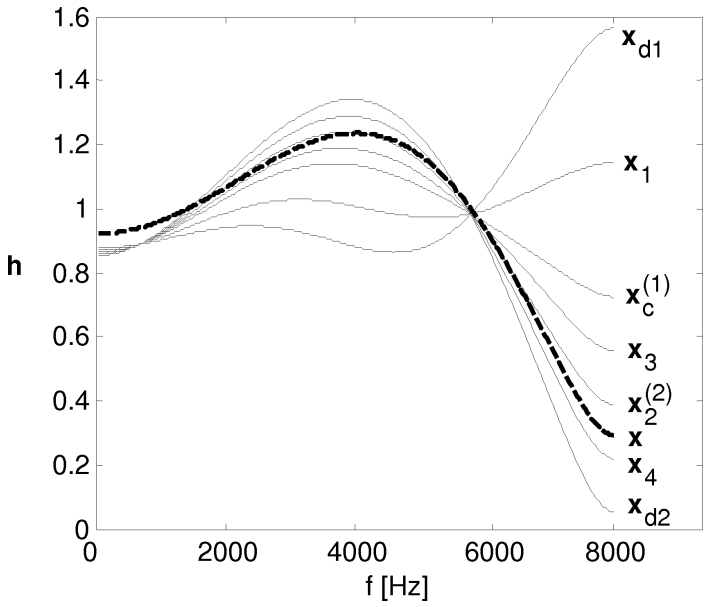

Fig. 7. Iterative determination ($J = 2$) of linear combination of elements $\mathbf{x}_c$, $\mathbf{x}_{d1}$ and $\mathbf{x}_{d2}$ closest to the given element $\mathbf{x}$. Combinations added in iteration 1 – $\mathbf{x}_1$ and $\mathbf{x}_2$; in iteration 2 – $\mathbf{x}_3$ and $\mathbf{x}_4$. The upper index denotes the closest combination in the given iteration.

8. Compute the total error in the i-th iteration $\varepsilon_i = ||(\mathbf{A}^{(r)} \cdot \mathbf{B}_i^{(r)} - \mathbf{X} + \mathbf{a}^{(r)}\mathbf{x}_c) \cdot \mathbf{L}||_F$, where index F stands for the Frobenius norm and matrix $\mathbf{L}$ was determined by optimization as described in Subsec. 5.1.

9. If $i < I$ and the decrease of error ε_i in relation to the error ε_{i-1} remains above the set threshold, return to step **2**.

10. Choose the 2nd and the 3rd rows of matrix $\mathbf{B}_i$ as additional elements $\mathbf{x}_{d1}$ and $\mathbf{x}_{d2}$ of the given cluster. In the case of filter clustering approximate the obtained frequency response magnitude by filter parameters of the given order.

The method used above does not guarantee the generally monotonic decrease of the error ε_i, however, in practice we observed a satisfactory decrease of this error. The number of iterations used in the implementation was $J = 4$ and $I = 8$. Figure 7 shows an example of the iterative computation of the linear combination of elements $\mathbf{x}_c$, $\mathbf{x}_{d1}$, $\mathbf{x}_{d2}$.

6. Assigning speakers to classes on the basis of F_0 estimation

The improvement of the recognition result by means of ST banks involves a high computational cost, as it means parallel recognition performed for many transformations. In order to partially reduce the computational effort required to assign the speakers into classes, F_0 estimation was applied. It was noted that in the case of two classes, the algorithms presented in Sec. 4 in principle divided speakers according to their sex.

The introduced algorithm of F_0 estimation is based on the amplitude spectrum of the frame, which is already determined in the process of MFCC parameterization. The method involves the detection of maxima and minima in the spectrum. The value of F_0 for the frame is computed as the weighed mean of distances between the maxima, while for utterances, it is computed as the mean of F_0 for frames whose weight values exceeded the preset threshold. The weights were defined in such a way so that they constituted the measure of the spectrum harmonicity.

The classification is performed by comparing the estimated F_0 value for utterances with the value of the decision threshold, which was set to 190 Hz on the basis of the data from the training set. The threshold value was determined in the way which corresponds to the Bayes decision criterion, which minimizes the mean decision error, provided that both speaker classes are equally probable [26].

7. Isolated words recognition

Isolated words recognition was performed by means of 4 ASR system versions. A vocabulary consisting of 18 words in Polish (10 digits and 8 commands) was used.

7.1. ASR system versions

Version A. Every phoneme was modeled by one HMM state. One transition probability matrix for the whole system was used. Modeling of phoneme duration probability

was applied. The first stage of utterance recognition involves finding the optimal path of states by means of the Viterbi algorithm followed by the division of this path into a sequence of pseudo-syllables. The second stage consists in comparing the obtained sequence of pseudo-syllables with the patterns of words from the vocabulary.

In the case of using the recognition improvement method by means of ST banks, the best solution was chosen with regard to the recognition score in the following representation:

$$P_{\text{sco}} = P_{\text{syl}} \cdot P_{\text{acou}}^{\gamma}, \tag{13}$$

where P_{syl} denotes pseudo-probability which is the measure of the adjustment of the recognized sequence of pseudo-syllables to the pattern of the given word. P_{acou} stands for the probability of the winning Viterbi path normalized by being raised to the $1/T$ power, where T denotes utterance duration in frames. Weight γ was experimentally set to 0.1.

Version At. Phonemes were modeled by means of two or three HMM states. The Baum-Welch algorithm was used in the training procedure [25]. No additional phoneme duration modeling was employed. In the Viterbi algorithm a constraint of preventing the obtaining of paths with forbidden three-phoneme sequences was used. Furthermore, a mechanism of analysing the best Q paths was introduced, where $Q = 8$ was obtained experimentally. The recognition score (13) was applied. The experimentally set weight in this version was $\gamma = 1.5$. The remaining elements of the system did not differ from those in version A.

Version B. The phoneme modeling in this version was the same as in version A. The phoneme duration probability modeling was also employed again. The recognition was performed by determining the probabilities of the winning Viterbi paths obtained from separate HMM models for each word from the vocabulary. Word models consisted of the sequences of states corresponding to phonemes in a given word. The path probability serves also as the recognition score when the recognition improvement method is used.

Version Bt. Phoneme models used in this version were similar to those used in version At. The modeling of the phoneme duration probability was not applied too. Other elements were the same as those in version B.

In recognition with employment of ST banks, the MFCC pdfs in classes are obtained according to the procedure described in Subsecs. 4.1–4.3. In versions At and Bt there are additional transition matrices and vectors of initial state probabilities for the phonemes. The values of these additional parameters are determined separately for each speaker class. The remaining elements of the system are the same for all classes. In the training procedure of a system designed to cooperate with the ST bank, we used a feedback consisting in cyclically performed recognition and choice of combinations of the speaker classes and ST bank elements combinations for each utterance, with the current system parameter values. These choices were taken into consideration in the further training. In the case of simulating transmission distortions, the system training procedure involves non-distorted data and ST banks that were not modified according

to the method presented in Subsec. 5.2. The recognition is performed by means of an appropriate modified bank.

7.2. Iterative improvement of recognition

The iterative recognition improvement algorithm outlined below is used for a speaker class and the ST bank elements chosen earlier for a given utterance on the basis of the recognition score analysis. As this score can be a discontinuous function, the following optimization method was selected. Moreover, the method is an optimization with constraints and ensures that the spectral transformation parameter values remain in the set limits.

1. Initialization. Set $j = 0$. Set matrix $\mathbf{B}_1$ to a matrix containing vectors $\mathbf{x}_c$, $\mathbf{x}_{d1}$ and $\mathbf{x}_{d2}$ of a given LF bank element in its subsequent rows (see Subsec. 5.3). Set matrix $\mathbf{B}_2$ to a matrix consisting of elements analogous to a given element of the $g(f)$ bank. Let matrix $\mathbf{A}_{1,0} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_2 \ \mathbf{e}_3 \ \mathbf{e}_3]^{\mathrm{T}}$, and matrix $\mathbf{A}_{2,0} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3 \ \mathbf{e}_2 \ \mathbf{e}_3]^{\mathrm{T}}$, where $\mathbf{e}_i$ denotes a 3-element column vector taking on the value of 1 on the i-th position and the value of 0 on other ones.
2. Increment j by 1. Compute the recognition score for $5 + 4(j-1)$ sets of spectral transformation parameter values obtained from the equations $\mathbf{E}_1 = \mathbf{A}_{1,j} \cdot \mathbf{B}_1$ and $\mathbf{E}_2 = \mathbf{A}_{2,j} \cdot \mathbf{B}_2$, respectively. Each set of parameter values consists of filter parameters (frequency response magnitude) and $g(f)$ parameters corresponding to the rows of matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ with the same number. Determine the row number of matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ (identical for both) for which the maximal recognition score was obtained.
3. Determine two row numbers of matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ (both numbers refer to the same row in the two matrices) for which a maximal recognition score was obtained, but for which the corresponding rows of matrix $\mathbf{A}_{1,j}$ differ from the row of matrix $\mathbf{A}_{1,j}$ with the number determined in step **2**.
4. Determine two subsequent row numbers of matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ for which the maximal recognition score was obtained, but for which the corresponding rows of matrix $\mathbf{A}_{2,j}$ differ from the row of matrix $\mathbf{A}_{2,j}$ with the number obtained in step **2**. Moreover, these are not row numbers chosen in step **3**.
5. If there were no 4 row numbers chosen in steps **3** and **4**, complete the row numbers to 4 by those, for which the maximal recognition score was obtained and which were not selected in steps **2–4**.
6. Create matrices $\mathbf{A}_{1,j+1}$ and $\mathbf{A}_{2,j+1}$ by adding 4 rows to matrices $\mathbf{A}_{1,j}$ and $\mathbf{A}_{2,j}$ that correspond to the means of the rows of matrices $\mathbf{A}_{1,j}$ and $\mathbf{A}_{2,j}$ with the number selected in step **2** and to the row numbers selected in steps **3–5**. This operation means the obtaining of 4 new points of direct search.
7. If $j < J$, return to step **2**. If this is not the case, compute the recognition score value for $5 + 4j$ sets of spectral transformation parameter values obtained from equations $\mathbf{E}_1 = \mathbf{A}_{1,j+1} \cdot \mathbf{B}_1$ and $\mathbf{E}_2 = \mathbf{A}_{2,j+1} \cdot \mathbf{B}_2$, respectively. Choose the result providing the maximal recognition score value as the final one.

The number of iterations in the implementation was set to $J = 4$.

8. Recognition results

Table 1 gives the numbers of the ST bank construction methods and the methods of speaker classes determination, which are further used in Tables 2 and 3. Table 2 presents isolated words recognition results for 4 versions of the ASR system in the case when there are no simulated transmission distortions. Table 3 shows the results obtained with the simulation of transmission distortions. The distortions were applied at random, separately for each utterance, first as one of the characteristics $\mathbf{q}_0$ to $\mathbf{q}_5$, and then as one of the characteristics $\mathbf{q}_0$ or $\mathbf{q}_6$ or $\mathbf{q}_7$ (see Subsec. 5.2). In all the cases, the results presented were obtained by means of a method guaranteeing the highest recognition. The changes in the recognition results after assigning speakers to classes by means of the F_0 estimation are provided in parentheses (columns 4–7). Recognitions without correction were performed by means of the system trained without dividing speakers into classes and without taking into consideration the spectral transformation (column 3).

Table 1. Numbers of ST banks construction and speaker classes determination methods used in Tables 2 and 3.

ST bank constr. method no.	ver. no. of speaker classes determ. method	ver. of function w used in (2)	ver. no. of value C_{conb} determ. method	Does the bank allow for transm. distort.?
1	2	w_1	1	no
2	3	w_2	2	no
3	3	w_1	1	no
4	3	w_1	2	no
5	1	w_2	1	yes
6	3	w_2	2	yes

Table 2. Isolated words recognition results obtained with no simulated transmission distortions.

speaker set	system version	result w/o correction [%]	correction w/o iterations		correction with iterations	
			bank constr. method no.	result [%]	bank constr. method no.	result [%]
1	2	3	4	5	6	7
training	A	98.00	4 (2)	98.67 (0.00)	4 (1)	98.89 (−0.44)
	At	98.89	4 (2)	99.33 (0.00)	4 (2)	99.33 (0.00)
	B	99.33	2 (2)	100.0 (0.00)	2 (2)	99.78 (+0.22)
	Bt	99.56	4 (2)	99.56 (+0.22)	4 (4)	99.56 (+0.22)
test	A	94.17	4 (4)	95.56 (−0.28)	2 (2)	96.11 (+0.56)
	At	95.28	2 (2)	97.22 (−1.39)	4 (1)	97.22 (0.00)
	B	96.11	2 (2)	97.78 (−0.28)	2 (3)	98.06 (−0.28)
	Bt	97.22	4 (2)	97.50 (+0.28)	2 (2)	97.78 (−0.28)

Table 3. Isolated words recognition results obtained with simulated transmission distortions.

speaker set	system version	result w/o correction [%]	correction w/o iterations		correction with iterations	
			bank constr. method no.	result [%]	bank constr. method no.	result [%]
1	2	3	4	5	6	7
training	A	92.44	5 (6)	97.33 (+0.67)	6 (6)	97.78 (+0.22)
	At	96.00	5 (5)	99.56 (−0.44)	5 (5)	99.33 (0.00)
	B	98.00	6 (5)	99.56 (−0.44)	6 (6)	99.56 (−0.22)
	Bt	97.11	5 (5)	99.56 (+0.22)	6 (5)	99.56 (0.00)
test	A	88.89	3 (6)	94.44 (0.00)	6 (6)	94.72 (+2.22)
	At	91.11	5 (5)	96.39 (+0.28)	6 (5)	96.94 (−1.11)
	B	94.72	6 (5)	96.94 (+0.84)	6 (6)	97.22 (0.00)
	Bt	96.67	6 (5)	97.78 (−0.28)	6 (6)	97.50 (−0.28)

Depending on the system version, the reduction of the recognition error rate with no simulated microphone distortions for the training set (the better result was selected from columns 5 and 7) varied from 0% (version Bt) to 100% (version B), and on average it amounted to 46%. For the test set it varied from 20% (version Bt) to 50% (version B), and on average it amounted to 36%. With the simulation of microphone distortions, the recognition error reductions varied for the training set from 71% (version A) to 89% (version At) and on average they amounted to 81%. For the test set they varied from 33% (version Bt) to 66% (version At) and on average they amounted to 50%.

The application of the iterative recognition improvement proved successful in 9 out of 16 cases presented in Tables 2 and 3 (column 7). The recognition turned out worse only in 2 cases. The majority (7 out 9) improvements were recorded in test sets after the iterative mechanism had been applied.

The application of the class selection on the basis of F_0 estimation had a neutral influence on the recognition results. In relation to the standard method of class determination, the change in recognition results, averaged with the data from Tables 2 and 3 (columns 5 and 7), amounted to $+0.007\%$ with the standard deviation of 0.597%.

We have not presented here results obtained for CMN, VTS, Eigenvoices, MLLR and cepstral parameters histogram equalization methods, which have been obtained in the preliminary research because these methods failed in the task of very short utterances recognition, i.e. they worsened the recognition results.

9. Conclusions

The paper presents a new method of linear transmission distortions compensation and speaker individual characteristics normalization designed to cooperate with very short isolated utterances recognition systems. The suggested approach employs spectral transformation banks and the division of speakers into classes. Other methods discussed

herein are the method of further recognition improvement by means of the iterative algorithm and the reduction of the computational complexity of the method by a preliminary selection of speaker classes on the basis the speaker's fundamental frequency.

The proposed solution has turned out efficient: even with a small training set, the average error rate reduction for various ASR system versions was 36% in the test set with no transmission distortions. With the simulation of these distortions it amounted to 50%. The highest recognition results were achieved mostly with the application of version 3 of the proposed speaker class determination method.

References

[1] ACERO A., *Acoustical and environmental robustness in automatic speech recognition*, PhD Thesis, Carnegie Mellon University, Department of Electrical and Computer Engineering, Pittsburgh 1990.

[2] MORENO P. J., RAJ B., STERN R. M., *A Vector Taylor series approach for environment-independent speech recognition*, Proc. International Conference on Acoustics, Speech and Signal Processing, Atlanta, USA 1996.

[3] DE LA TORRE A., PEINADO A. M., SEGURA J. C., PÉREZ-CÓRDOBA J. L., BENÍTEZ M. C., RUBIO A. J., *Histogram equalization of speech representation for robust speech recognition*, IEEE Trans. on Speech and Audio Processing, **13**, 3, 355–365 (2005).

[4] RAJ B., SELTZER M. L., STERN R. M., *Reconstruction of missing features for robust speech recognition*, Speech Communication, **43**, 275–296 (2004).

[5] ZHU D., NAKAMURA S., PALIWAL K. K., WANG R., *Maximum likelihood sub-band adaptation for robust speech recognition*, Speech Communication, **47**, 243–264 (2005).

[6] MING J., *Noise compensation for speech recognition with arbitrary additive noise*, IEEE Trans. on Audio, Speech and Language Processing, **14**, 3, 833–844 (2006).

[7] LEE L., ROSE R., Member, *A frequency warping approach to speaker normalization*, IEEE Trans. on Speech and Audio Processing, **6**, 1, 49–60 (1998).

[8] MCDONOUGH J., SCHAAF T., WAIBEL A., *Speaker adaptation with all-pass transforms*, Speech Communication, **42**, 75–91 (2004).

[9] GUNAWARDANA A., BYRNE W., *Discriminative speaker adaptation with conditional maximum likelihood linear regression*, Proc. Eurospeech, Aalborg, Denmark 2001.

[10] CHEN K.-T., WANG H.-M., *Eigenspace-based linear transformation approach for rapid speaker adaptation*, Proc. ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition, Sophia Antipolis, France 2001.

[11] SHINODA K., LEE C.-H., *A structural Bayes approach to speaker adaptation*, IEEE Trans. on Speech and Audio Processing, **9**, 3, 276–287 (2001).

[12] KIM D. K., KIM N. S., *Maximum a posteriori adaptation of HMM parameters based od speaker space projection*, Speech Communication, **42**, 59–73 (2004).

[13] GALES M. J. F., *Cluster adaptive training of Hidden Markov Models*, IEEE Trans. on Speech and Audio Processing, **8**, 4, 417–428 (2000).

[14] KUHN R., JUNQUA J.-C., NGUYEN P., NIEDZIELSKI N., *Rapid speaker adaptation in eigenvoice space*, IEEE Trans. on Speech and Audio Processing, **8**, 6, 695–707 (2000).

[15] MAK B., KWOK J. T., HO S., *Kernel eigenvoice speaker adaptation*, IEEE Trans. on Speech and Audio Processing, **13**, 5, 984–992 (2005).

[16] EVEREST F. A., *The master handbook of acoustics*, McGraw-Hill, 2001.

[17] GROCHOLEWSKI S., *First database for spoken Polish*, Proc. International Conference on Language Resources and Evaluation, Grenada 1998.

[18] MRÓWKA P., MAKOWSKI R., *Channel and speaker variety compensation using modified eigenvoices algorithm*, Proc. International Conference on Signals and Electronic Systems, Poznań, Poland 2004.

[19] SPALL J. C., *Adaptive stochastic approximation by the simultaneous perturbation method*, IEEE Trans. on Automatic Control, **45**, 10, 1839–1853 (2000).

[20] ARABAS J., *Lectures on evolutionary algorithms* [in Polish], WNT, Warszawa, 2004.

[21] CHELOUAH R., SIARRY P., *Genetic and Nelder–Mead algorithms hybridized for a more accurate global optimization of continuous multiminima functions*, European Journal of Operational Research, **148**, 335–348 (2003).

[22] MAK B., BARNARD E., *Phone clustering using the Bhattacharyya distance*, Proc. International Conference on Spoken Language Processing, Philadelphia, USA 1996.

[23] KAILATH T., *The divergence and Bhattacharyya distance measures in signal selection*, IEEE Trans. on Communication Technology, **15**, 1, 52–60 (1967).

[24] FLETCHER R., *Practical methods of optimization*, John Wiley & Sons, 1980.

[25] RABINER L., JUANG B. H., *Fundamentals of speech recognition*, Prentice Hall, New Jersey 1993.

[26] FRANKS L. E., *Signal theory* [in Polish], PWN, Warszawa, 1975.