www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. 41, No. 1, pp. 107–118 (2016)

DOI: 10.1515/aoa-2016-0011

# An Effective Speaker Clustering Method using UBM and Ultra-Short Training Utterances

Robert HOSSA, Ryszard MAKOWSKI

*Signal Processing Systems Department*
*Faculty of Electronics*
*Wroclaw University of Technology*
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland; e-mail: {robert.hossa, ryszard.makowski}@pwr.edu.pl

(*received November 13, 2014; accepted November 8, 2015*)

The same speech sounds (phones) produced by different speakers can sometimes exhibit significant differences. Therefore, it is essential to use algorithms compensating these differences in ASR systems. Speaker clustering is an attractive solution to the compensation problem, as it does not require long utterances or high computational effort at the recognition stage. The report proposes a clustering method based solely on adaptation of UBM model weights. This solution has turned out to be effective even when using a very short utterance. The obtained improvement of frame recognition quality measured by means of frame error rate is over 5%.

It is noteworthy that this improvement concerns all vowels, even though the clustering discussed in this report was based only on the phoneme a. This indicates a strong correlation between the articulation of different vowels, which is probably related to the size of the vocal tract.

**Keywords:** automatic speech recognition; interindividual difference compensation; speaker clustering; universal background model; GMM weighting factor adaptation.

## 1. Introduction

While building an automatic system of speech recognition (ASR), one of the first steps is to work out an acoustic model of the analysed language. Devising such a model usually requires: (i) partial or complete segmentation and labelling of training set recordings, (ii) selection of a parameterisation method and its implementation (iii) approximation of probability distribution estimators of parameter (observation vector) values for particular phonemes by the sum of Gaussian distributions – a GMM model.

The same speech sounds produced by different speakers are sometimes very different. What ensues is the flattening of GMM model probability distribution and, consequently, deterioration of its classification abilities. These differences are referred to as interindividual differences and they are caused by differences in the speakers' anatomy (e.g. the size of the vocal tract) as well as their different personalities. Although other factors such as contextual differences, environmental conditions, or intraindividual differences also have an impact on parameter values, interindividual differences are among the crucial ones. There are various differences in the time-frequency structure of speech sounds (phones), including divergences in the frequency and width of particular formants, which are reflected in the values of the observation vector. In terms of statistics, these differences are very clear between members of opposite sexes and children. For instance, formant frequencies for female voices are on average about 17% higher than those for male voices (Jassem, 1973).

Losses in the classification abilities of the acoustic model caused by the mentioned factors can be reduced by using additional algorithms referred to as compensation algorithms. These algorithms are used at different stages of recognition in a variety of ways: (i) robust parameterisation algorithms are used at parameterisation stage, (ii) after parameterisation, modification (usually standardisation) of the observation vector is employed, (iii) speaker clustering is applied at ASR system training stage, and (iv) adaptive modification of a statistical GMM model is used at the stage of determination of phonetic unit sequences (Makowski, 2011). For various reasons, not all of these algorithms

can be used in all conditions. In all compensation algorithms, the key role is played by the length of the utterance which can be used. If the utterance is intended to be short, employing complicated compensation methods, which usually involve estimation of many parameters, is out of the question. A general rule is applied, saying that the greater number of parameters needs to be estimated, the longer the utterance used for this purpose has to be.

The presented compensation model is one of the key problems related to designing ASR systems and the literature on the subject is very rich (e.g. Kuhn *et al.*, 2000; Hazen, 2000; Naito *et al.*, 2002; Mak *et al.*, 2004; de la Torre *et al.*, 2005; Mrówka, Makowski, 2007). Studies and analyses point to a possibility of at least partial compensation of the unfavourable influence that the mentioned factors have on the recognition quality.

The present report will focus on speaker clustering, which consists of building separate statistical acoustic models for speaker groups or conditions characterised by similar features. The criterion of such a division can be based on: the frequency of pitch correlated with the speaker's sex, the level of interference, observation vectors, etc. As the application of clustering at the recognition stage does not require long utterances or long and complicated computations, it is an attractive solution. While embarking on speaker clustering, one has to compromise two contradictory aspects: (i) the larger number of groups, the better adjustment of the model to the specific features of a particular group, so the model will be more adequate for that group, (ii) the larger number of groups, the lower the accuracy of the model, since fewer speakers will be available for each model.

Naito *et al.* (2002) proposed a clustering method based on vocal tract parameters extracted from the signal. Kosaka and Sagayama (1994) proposed the hierarchical clustering (HC) based on probability distances obtained from hidden Markov networks. On the very top of this structure, there is a model encompassing all the speakers, while individual models can be found at the very bottom. In the recent years, speaker clustering has often targeted a similar problem of identifying signal segments coming from a given speaker and designating them jointly with a clear-cut term (speaker diarisation problem) (e.g., Tranter, Reynolds, 2006). Such a problem, alongside preliminary segmentation of speech signal, is the key element of creating a speakers' dictionary. Owing to the lack of knowledge regarding the real number of speakers, the discussed problem can be interpreted as an unsupervised learning problem. Classic solutions to the problem formulated in this way are based on the concept of hierarchical clustering, where different numbers of groups are checked and various distance measures are used: Mahalanobis distance (Iyer *et al.*, 2006), Bhattacharyya distance (Basseville, 1989), Hellinger distance (Lu *et al.*, 2003), and Generalized Likelihood Ratio (GLR) (Anderson, 2003). Further development of the discussed HC methods has given rise to the clustering algorithm applying Leader Following Concept (LFC) to k-means algorithm (Duda *et al.*, 2001). It was followed by a hybrid algorithm based on GLR and Global Dispersion Criterion devised by Liu and Kubala (2004). Other alternative solutions of the speaker clustering problem use the concept of Universal Background Model (UBM) of low level acoustic feature vectors (Reynolds, Rose, 1995; Tang *et al.*, 2012) based on the Gaussian Mixture Model (GMM) and they require a training stage with a large number of recordings. The problem of access to such a large number of data can be partially solved by employing training techniques based on the Maximum a Posteriori (MAP) method. In the recent years, one can also observe an increased interest in applying the concept of supervectors, incorporating mean vectors of all the GMM mixture components, to speaker clustering and in effective methods of reducing the dimensionality of these supervectors by means of PCA (Principal Component Analysis) (Bishop, 2006), LPP (Locally Preserving Projection) (He, Niyogi, 2003), or LDA (Linear Discriminant Analysis) algorithms (Chu *et al.*, 2009; Mehrabani, Hansen, 2013). Another rapidly developing group of speaker clustering solutions uses the BIC (Bayesian Information Criterion) (Stafylakis *et al.*, 2006; Tsai *et al.*, 2007), which is a measure of distance between two statistical models.

## 2. Problem formulation

The aim of the presented research is an optimal division of speakers into groups in the meaning defined in Sec. 3. Let us assume that the number of groups $G$ is known and it is not a very large number. We assume that clustering and then assigning a speaker to a particular group (cluster) are both based on a very short utterance, which is a fragment comprising the vowel *a* from a single utterance of the word *tak* (*yes*). This is the basic variant designated as 1. For comparison purposes, the division is also performed in two other variants: variant 2, based on pieces of various words containing the phoneme *a* and variant 3, based on pieces of various words containing any vowel. The first approach is justified in command recognition systems, where it is easy to enforce the answer *tak (yes)* at the beginning of a recognition session, and then, by using automatic segmentation (e.g., Makowski, Hossa, 2014), extract the vowel *a*. After appropriate modification, such an approach could be also used in continuous speech recognition systems to decide which acoustic model should be used in the case of a speaker change. An advantage of such an approach would be the absence of the requirement of possessing a long

piece of utterance at the stage of assigning the speaker to a cluster. Variant 2 aims at assessing the representativeness of approach 1 for clustering based on a larger number of observations with different phonetic contexts. Finally, variant 3 assesses to what extent *a*-based clustering is representative of all vowels, since reliable recognition of all vowels is of prominent importance for the effectiveness of all the ASR systems.

In order to assess the effectiveness of the various clustering methods, having a set of segmented and labelled recordings, one should: (i) cluster the speakers of the training set based on the set of observation vectors, (ii) generate an acoustic model for each speaker cluster based on all the training set recordings, and (iii) assess the improvement of recognition quality for a particular clustering method used for the training and the checking set, with the use of a predefined measure.

## 3. Speaker clustering methods

### *3.1. Proposal of UBM-based method – method 1*

#### *3.1.1. Universal background model*

Let us assume that we have access to recordings of $P$ speakers, $P \gg G$, producing short utterances. The first and basic method of speaker clustering is based on the UBM technique (Reynolds et al., 2000; Chu *et al.*, 2009a; 2009b; Mehrabani, Hansen, 2013). A flowchart illustrating the proposed method is shown in Fig. 1. It is valid for all the three variants of clustering, but different words from the database are used in particular variants (the words *tak* for variant, etc.).

From the speakers' utterances, observation vectors (low level acoustic feature vectors) are extracted, e.g. MFCC, for which a GMM statistical model, being a mixture of Gaussian distributions, is built. Thus, in the large observation set comprising all the speakers, an acoustic model is determined for the phoneme *a* (in experiments 1 and 2) or for $F$ selected phonemes (e.g., for all the vowels of the Polish language) in variant 3 by using an EM (Expectation Maximisation) algorithm. In this way, we create universal background models UBM, whose parameters form a baseline model for adaptive MAP estimation methods. Every UBM model is a mixture composed of $G$ normal distributions described by: mean value vectors $\{\mathbf{m}_1, \mathbf{m}_2, ...\mathbf{m}_G\}$, full covariance matrices $\{\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_G\}$, and weighting factors $\{w_1, w_2, ..., w_G\}$, i.e.

$$p_U(\mathbf{o}) = \sum_{i=1}^{G} w_i \mathcal{N}(\mathbf{o}, \mathbf{m}_i, \mathbf{R}_i), \tag{1}$$

where

$$\sum_{i=1}^{G} w_i = 1, \tag{2}$$

$$\mathcal{N}(\mathbf{o}, \mathbf{m}_i, \mathbf{R}_i) = \frac{1}{(2\pi \det(\mathbf{R}_i))^{M/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{o}-\mathbf{m}_i)^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{o}-\mathbf{m}_i)\right) \tag{3}$$

and $\mathbf{o}$ is the observation vector, $M$ is its length.

#### *3.1.2. Model adaptation*

Based on the utterance or utterances relevant to a particular variant, the UBM model is subject to MAP adaptation for every speaker separately. In the proposed solution, only the weighting factors of UBM model are changed. Their interpretation of *a priori* probabilities is that a given observation set $\mathbf{o}_t$ belongs to the $i$-th acoustic class described by distribution (3). In this way, we expect to adapt the acoustic model to a particular speaker based solely on the vowel *a*. With the assumption that the prior distribution of weighting factors is a multidimensional Dirichlet distribution with linear concentration parameters, the MAP formula of weight re-estimation assumes the form:

$$w_{pi}^{(')} = \frac{\eta_{pi}}{T_p} \tag{4}$$

or alternatively (Reynolds *et al.*, 2000; Chu *et al.*, 2009a; 2009b)

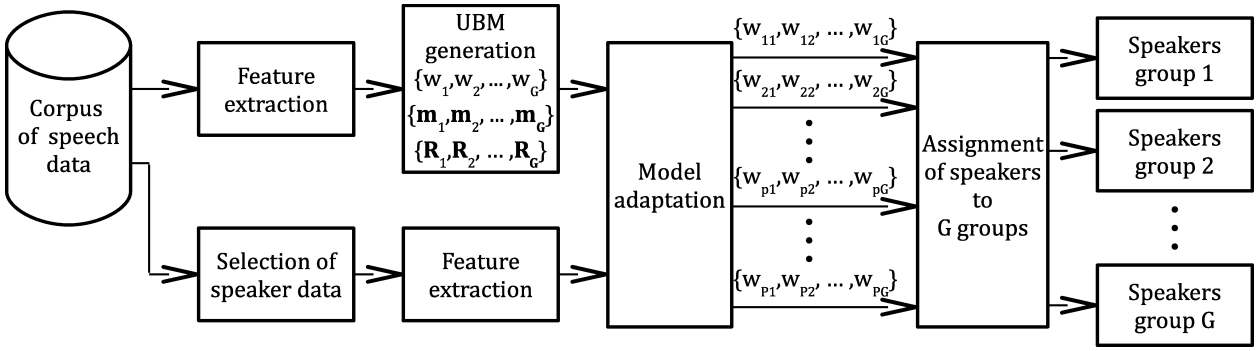


Fig. 1. Flowchart showing UBM-based speaker clustering.

$$w_{pi}^{(')} = \left[\alpha_i \frac{\eta_{pi}}{T_p} + (1-\alpha_i) w_{pi}\right] \delta, \tag{5}$$

where $p$ is the speaker number, $w_{pi}^{(')}$ is the new weight value, $T_p$ is the number of speech signal frames which form the basis for adaptation, $\alpha_i$ is the adaptation constant, and $\delta$ is a scaling factor securing the satisfying of relation (2). Then, $\eta_{pi}$ is the sum of probabilities produced by the following relation:

$$\eta_{pi} = \sum_{t=1}^{T_p} p(i|\mathbf{o}_t), \tag{6}$$

where $p(i|\mathbf{o}_t)$ is the probability of belonging of observation $\mathbf{o}_t$ to group $i$, according to Bayes' rule, by the relation:

$$p(i|\mathbf{o}_t) = \frac{w_{pi}\mathcal{N}(\mathbf{o}_t|\mathbf{m}_i, \mathbf{R}_i)}{\sum_{i=1}^{G} w_{pi}\mathcal{N}(\mathbf{o}_t|\mathbf{m}_i, \mathbf{R}_i)}. \tag{7}$$

The adaptation described by (4)–(7) is repeated until the relative change in the value of factor $w_{pi}$ exceeds the assumed threshold $\lambda$. Experiments have shown that adaptations (4) and (5) lead to very similar results and differences are due to a slightly different method of reaching the assumed threshold $\lambda$. At this point, it is worth emphasising that the advantage of adaptation rule (4) is reaching the threshold many times faster and requiring much fewer computations as compared to rule (5).

#### *3.1.3. Arrangement of the experiments*

After the completion of the adaptation process, the speakers from the training set are assigned to one of $G$ groups according to the weight values obtained through adaptation $\{w_{pi} : p = 1, ..., P; i = 1, ..., G\}$. In the same way, the speakers from the checking set are assigned to groups. Having access to the speakers divided into groups, we can determine $G$ sets of acoustic models (separately for each speaker group) comprising all the phonemes and a set of common models for all the speakers. Depending on the experiment variant, these models differ from each other. As a result, having such acoustic models at disposal, we can assess the effectiveness of the employed methodology of speaker clustering. A flowchart presenting the proposed methods of analysing this effectiveness is shown in Fig. 2.

### *3.2. Minimisation of distance L2 between group members and group means – method 2*

The traditional *k-means* method was used as a reference method here. In this method, division of speakers into groups is performed on the principle of minimising the mean distances of particular speakers' observation vectors from the mean observation vectors for the group. The starting point for clustering is an arbitrary preliminary division of speakers into $G$ groups and the division algorithm assumes the form:

1. Determining mean values $\mathbf{m}_p$ of observation vector coefficients for particular speakers and sub-

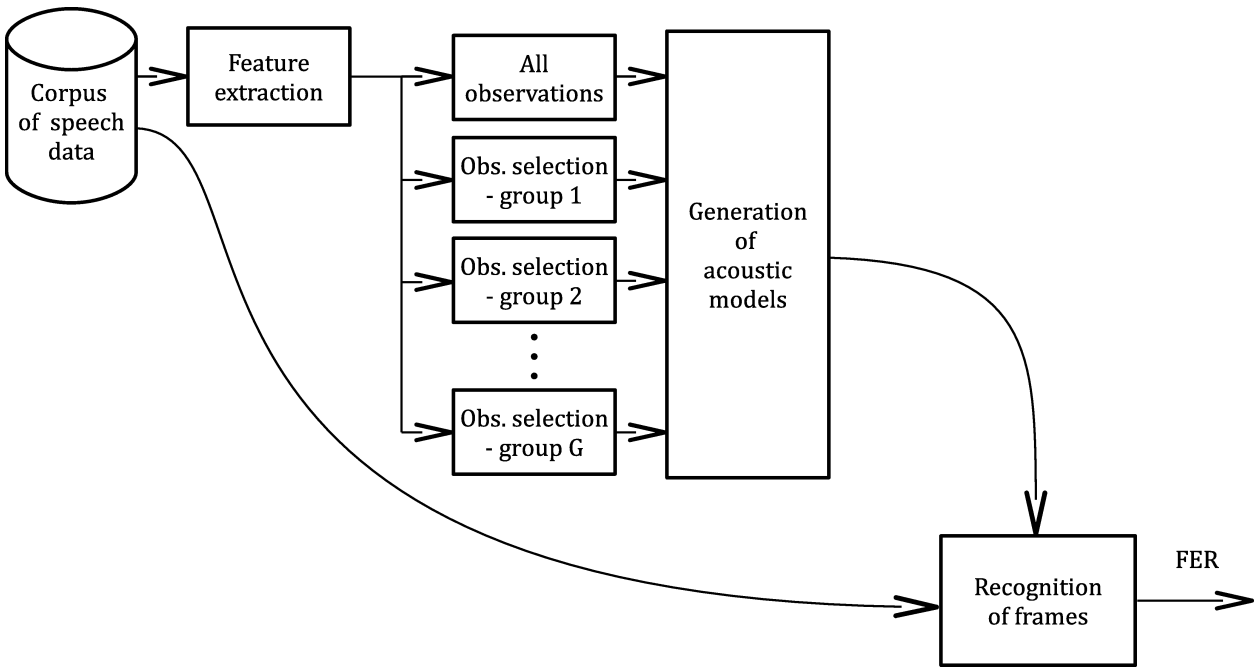


Fig. 2. Defining acoustic model groups and recognition quality measures (FER).

sequent coefficients, in accordance with the relation:

$$\mathbf{m}_p = \frac{1}{T_p}\sum_{t=1}^{T_p}\mathbf{o}_{tp}, \tag{8}$$

where $p$ is the speaker number and $T_p$ is the number of frames of included observations.

2. Determining the mean vector $\mathbf{m}_g$ each group:

$$\mathbf{m}_g = \frac{1}{N_g}\sum_{i=1}^{N_g}\mathbf{m}_i; \qquad g = 1, 2, ..., G, \tag{9}$$

where $N_g$ is the number of speakers in a group.

3. Calculation of distance $d_{pg}^{(2)}$ of each speaker's mean vector $\mathbf{m}_p$ from the mean vector $\mathbf{m}_g$ of particular groups:

$$d_{pg}^{(2)} = \sqrt{\sum_{m=1}^{M}\left(m_p(m) - m_g(m)\right)^2}, \tag{10}$$

where $m$ is the index of observation vector coefficient.

4. If for any speaker from any group $g$ the distance $d_{pg}^{(2)}$ from their own group is longer than the distance from another group, then the speaker is transferred to the group lying within the shortest distance $d_{pg}^{(2)}$, which is followed by a jump to point 2 of this algorithm. Otherwise, the algorithm ends.

As experience shows, division into groups is strongly dependent on the initial grouping conditions. Therefore, the speaker assignment procedures described below were performed repeatedly for different preliminary divisions into groups. In this way, we obtained many divisions into groups which satisfied the criterion of minimising intragroup distances. These solutions were assessed based on deflection coefficients and the best solution was chosen. The criterion of the best solution choice was the maximisation of the coefficient defined by the following relation:

$$U = \sum_{i,j} U_{ij}; \qquad i, j = 1, ..., G; \qquad i > j, \tag{11}$$

where $U_{ij}$ is the deflection coefficient produced by the relation:

$$U_{ij} = \sum_{m=1}^{M}\frac{\left[m_i(m) - m_j(m)\right]^2}{\sigma_i(m)\sigma_j(m)}. \tag{12}$$

Relation (12) implies that the longer the distance between the means $m_i(m)$ and $m_j(m)$ and the smaller the variances $\sigma_i^2(m)$ and $\sigma_j^2(m)$, the higher the value of $U_{ij}$. The coefficient $U_{ij}$ expresses numerically the law on classification abilities of probability distributions, known from detection theory.

### 3.3. Clustering quality measures

Clustering quality can be assessed in many ways. The most reliable one would be assessing the recognition quality of all the ASR systems. However, such an approach requires employing many recognition levels and multilevel training. Therefore, simpler measures such as word error rate (WER) or frame error rate (FER) are preferable. The recognition quality measure used in this report is the FER defined by the relation:

$$\mathrm{FER} = \frac{T_{\mathrm{err}}}{T_c}\cdot 100\%, \tag{13}$$

where $T_{\mathrm{err}}$ is the number of misrecognised frames and $T_c$ is the number of all the analysed frames.

## 4. Clustering effectiveness results

### 4.1. Recording database and acoustic models

The set of recordings, being the experiment database, comprises recordings of 36 adult male voices registered in various Polish cities. 150 Polish words were recorded for each speaker. The recording database was divided into the training set – 24 speakers and the test set – 12 speakers. This division is represented in column 2 of Table 1. The signal sampling frequency was 12 kHz. The presented results refer to noisy signals with the signal/noise ratio of 30 dB. The phonetic description of speech was based on the set of 37 phonemes proposed by Jassem (1973), supplemented with 1 speech sound connected chiefly with transitional stages, especially near pauses in utterances. All these recordings were subjected to hand segmentation and labelling.

The MFCC method was used as the parameterisation method. The frame length was 20 ms and the frame step was 10 ms. The set of 14 cepstral coefficients was supplemented with an energy index and their first and second derivatives. Consequently, the length of the observation vector was 45. The acoustic models used at the frame recognition stage (cf. Fig. 2) are a mixture of $K = 5$ multidimensional normal probability distributions with a diagonal covariance matrix, i.e.

$$p_f(\mathbf{o}) = \sum_{i=1}^{K} w_{fi}\mathcal{N}(\mathbf{o}, \mathbf{m}_{fi}, \mathbf{\Sigma_{fi}}), \tag{14}$$

where

$$\mathbf{\Sigma_{fi}} = \begin{bmatrix} \sigma_{fi1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{fi2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{fiM}^2 \end{bmatrix} \tag{15}$$

Table 1. Values of GMM weights adapted with the use of observations of the phoneme *a* from the word *tak* (columns 3–5), *a* from many different words (columns 6–8) and the averaged weights for vowels (columns 9–11).

| speaker | set | variant 1 | | | variant 2 | | | variant 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $w_{p1}$ | $w_{p2}$ | $w_{p3}$ | $w_{p1}$ | $w_{p2}$ | $w_{p3}$ | $\overline{w}_{p1}$ | $\overline{w}_{p2}$ | $\overline{w}_{p3}$ |
| mbi02 | training | 0.80 | 0.20 | 0.00 | 0.47 | 0.33 | 0.20 | 0.50 | 0.30 | 0.20 |
| mbi04 | training | 0.00 | 0.00 | 1.00 | 0.72 | 0.14 | 0.14 | 0.47 | 0.19 | 0.34 |
| mbi05 | training | 0.00 | 1.00 | 0.00 | 0.00 | 0.79 | 0.21 | 0.07 | 0.66 | 0.27 |
| mka02 | test set | 0.89 | 0.11 | 0.00 | 0.66 | 0.33 | 0.01 | 0.34 | 0.57 | 0.09 |
| mka03 | training | 1.00 | 0.00 | 0.00 | 0.83 | 0.16 | 0.01 | 0.83 | 0.15 | 0.02 |
| mka04 | training | 0.33 | 0.00 | 0.67 | 0.22 | 0.16 | 0.62 | 0.46 | 0.17 | 0.37 |
| mle01 | training | 0.95 | 0.05 | 0.00 | 0.89 | 0.11 | 0.00 | 0.67 | 0.24 | 0.09 |
| mle06 | training | 0.00 | 1.00 | 0.00 | 0.06 | 0.85 | 0.09 | 0.25 | 0.35 | 0.40 |
| mle09 | training | 0.00 | 1.00 | 0.00 | 0.30 | 0.70 | 0.00 | 0.48 | 0.31 | 0.21 |
| mlu01 | test set | 0.06 | 0.19 | 0.75 | 0.01 | 0.09 | 0.90 | 0.08 | 0.25 | 0.67 |
| mlu04 | test set | 0.45 | 0.55 | 0.00 | 0.29 | 0.70 | 0.01 | 0.27 | 0.50 | 0.23 |
| mlu08 | training | 0.87 | 0.03 | 0.10 | 0.90 | 0.09 | 0.01 | 0.67 | 0.22 | 0.11 |
| mnt01 | test set | 0.00 | 0.31 | 0.69 | 0.12 | 0.12 | 0.76 | 0.05 | 0.34 | 0.62 |
| mnt05 | test set | 1.00 | 0.00 | 0.00 | 0.62 | 0.35 | 0.03 | 0.23 | 0.43 | 0.34 |
| mnt09 | training | 0.90 | 0.10 | 0.00 | 0.82 | 0.18 | 0.00 | 0.45 | 0.53 | 0.02 |
| mol03 | training | 0.00 | 1.00 | 0.00 | 0.14 | 0.86 | 0.00 | 0.24 | 0.62 | 0.14 |
| mol05 | test set | 0.61 | 0.39 | 0.00 | 0.31 | 0.39 | 0.30 | 0.14 | 0.71 | 0.15 |
| mol06 | training | 0.72 | 0.28 | 0.00 | 0.73 | 0.27 | 0.00 | 0.63 | 0.33 | 0.04 |
| mon03 | training | 0.13 | 0.00 | 0.87 | 0.22 | 0.14 | 0.63 | 0.48 | 0.28 | 0.24 |
| mon08 | training | 0.00 | 0.00 | 1.00 | 0.00 | 0.05 | 0.95 | 0.04 | 0.12 | 0.84 |
| mon09 | training | 0.00 | 1.00 | 0.00 | 0.15 | 0.82 | 0.03 | 0.19 | 0.66 | 0.15 |
| mry03 | test set | 0.86 | 0.02 | 0.12 | 0.36 | 0.23 | 0.41 | 0.20 | 0.66 | 0.14 |
| mry05 | training | 0.00 | 0.08 | 0.92 | 0.00 | 0.20 | 0.80 | 0.00 | 0.24 | 0.76 |
| mry09 | training | 0.00 | 0.00 | 1.00 | 0.00 | 0.13 | 0.87 | 0.25 | 0.36 | 0.39 |
| mrz06 | test set | 0.00 | 1.00 | 0.00 | 0.11 | 0.86 | 0.03 | 0.10 | 0.56 | 0.34 |
| mrz07 | test set | 0.90 | 0.00 | 0.10 | 0.94 | 0.04 | 0.02 | 0.74 | 0.10 | 0.16 |
| mrz10 | training | 0.00 | 0.00 | 1.00 | 0.00 | 0.07 | 0.93 | 0.06 | 0.12 | 0.82 |
| mwa02 | training | 0.78 | 0.11 | 0.11 | 0.75 | 0.11 | 0.14 | 0.58 | 0.25 | 0.17 |
| mwa04 | training | 0.00 | 0.03 | 0.97 | 0.00 | 0.19 | 0.81 | 0.01 | 0.16 | 0.83 |
| mwa09 | training | 0.00 | 1.00 | 0.00 | 0.00 | 0.84 | 0.16 | 0.01 | 0.24 | 0.75 |
| mwi03 | test set | 0.24 | 0.00 | 0.76 | 0.02 | 0.07 | 0.91 | 0.21 | 0.45 | 0.34 |
| mwi06 | training | 0.00 | 0.00 | 1.00 | 0.03 | 0.13 | 0.84 | 0.02 | 0.37 | 0.61 |
| mzw04 | training | 0.00 | 0.15 | 0.85 | 0.52 | 0.18 | 0.30 | 0.16 | 0.34 | 0.50 |
| mzw05 | test set | 0.00 | 0.86 | 0.14 | 0.74 | 0.26 | 0.00 | 0.62 | 0.33 | 0.05 |
| mzw10 | training | 1.00 | 0.00 | 0.00 | 0.95 | 0.04 | 0.01 | 0.76 | 0.23 | 0.01 |
| mwr31 | test set | 1.00 | 0.00 | 0.00 | 0.57 | 0.43 | 0.00 | 0.61 | 0.28 | 0.11 |

and

$$\mathcal{N}(\mathbf{o}, \mathbf{m}_{fi}, \mathbf{\Sigma}_{fi}) = \prod_{m=1}^{M} \frac{1}{\sqrt{(2\pi)}\sigma_{fi}(m)} \cdot e^{-\frac{1}{2\sigma_{fi}^2(m)}[o(m)-m_{fi}(m)]^2}. \quad (16)$$

It is generally known that the values of the observation vector for each phoneme are characterised by a high volatility. Figure 3 presents the mean values of MFCC features for successive coefficients and for frames building the phoneme *a* from the word *tak* for 3 example speaker groups (top left graph) and the mean

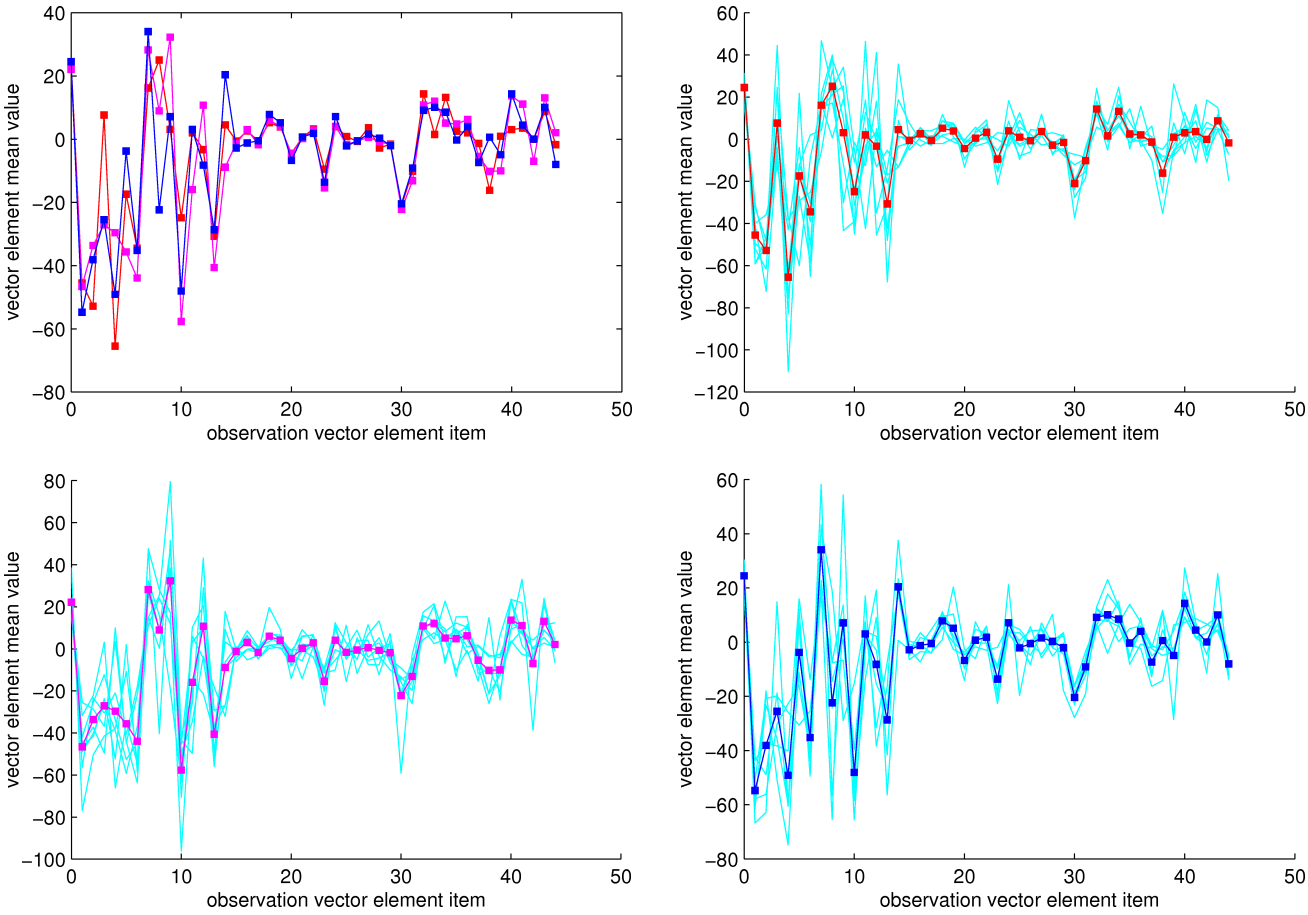


Fig. 3. Mean values of MFCC coefficients: means for 3 sets resulting from division into groups – top left graph, means for group speakers, and means for groups 1, 2, and 3 – the remaining graphs respectively.

values for speakers and means for groups 1, 2, and 3 (the remaining graphs respectively).

### *4.2. Variant 1 – clustering based on the phoneme* a *from the word* tak

#### *4.2.1. Cases 1 and 2*

For cases 1 and 2 of the experiment, the UBM for phoneme *a* was determined from many words coming from segmented recordings of all the 36 speakers. In this model, the weighting factors for $G = 3$ take the values $\{0.37; 0.31; 0.32\}$. The values of these weights are comparable, which means that the number of observations is similar for each component of sum (1). Subsequently, adaptation of UBM model weights was performed for each individual speaker, based on the observation of the phoneme *a* from the word *tak*. The number of these observations oscillated between 8 and 14. Let us denote these factors as $w_{pg}$, where $p$ is the speaker number and $g$ is the group number. The results of such adaptation are shown in columns 3–5 of Table 1. For most speakers, the observations clearly belong to one of the classes (the value of one of the weights is close to one).

In the next step, the speakers from the training set were divided into 3 groups based on $w_{pg}$ values. As a result, the number of speakers in set 1 reached 8, in set 2–6 and in set 3–10. As a consequence of this division, 4 acoustic models were built (one for the common and three for particular groups) for all the 38 states (phones). Based on all the segmented recordings, separate FER measures were determined for the training and the test sets by using one acoustic model, 3 models (for the training set), or 4 models (for the test set). The results of frame recognition effectiveness are presented in columns 3–5 of Table 2. The absence of a result means that it is the same as the one in the adjacent column on the left.

Column 3 (case 0) contains the results obtained without clustering, i.e., one common set of acoustic models was used for all the speakers. Column 4 (case 1), for the training set, comprises recognition results obtained when the sets of GMM models used for a given speaker had been determined for the group where the speaker belonged. For the test set, group model sets were used for those speakers for whom the highest value of weight $w_{pg}$ exceeded the value 0.9 (4 speakers), and common sets – for the remaining

Table 2. Values of FER as a recognition quality measure for the case without speaker clustering and for different variants of clustering, for the observation of the phoneme *a* from the word *tak*.

| phonemes | set | case 0 | case 1 | case 2 | case 3 | case 4 |
|---|---|---|---|---|---|---|
| i | training | 26.1% | 20.0% | | 20.2% | |
| i | test set | 26.0% | 26.3% | 27.6% | 26.3% | 26.4% |
| ɨ | training | 37.3% | 29.1% | | 27.0% | |
| ɨ | test set | 42.4% | 43.5% | 48.4% | 45.3% | 46.5% |
| e | training | 34.7% | 30.4% | | 31.2% | |
| e | test set | 41.1% | 41.7% | 43.2% | 40.9% | 39.7% |
| a | training | 20.2% | 16.6% | | 16.5% | |
| a | test set | 25.2% | 23.4% | 23.2% | 25.5% | 25.1% |
| o | training | 26.4% | 20.7% | | 20.8% | |
| o | test set | 32.7% | 31.6% | 33.0% | 34.3% | 34.1% |
| u | training | 41.6% | 32.8% | | 32.6% | |
| u | test set | 44.9% | 47.3% | 47.0% | 45.0% | 46.3% |
| vowels | training | 28.7% | 23.5% | | 23.5% | |
| vowels | test set | 33.6% | 33.3% | 34.5% | 34.3% | 34.0% |
| all | training | 33.1% | 27.7% | | 27.8% | |
| all | test set | 40.2% | 40.7% | 41.8% | 40.9% | 41.1% |

ones. In these cases, four model sets were used. Such a procedure is justifiable, since if all the $w_{pg}$ factors for a given speaker have a low value, it means that no group is representative enough of them and they are better represented by the averaged set. Column 5 (case 2) comprises the results equivalent to those in column 4, and the threshold of group assignment is lowered to 0.85. Consequently, the number of speakers assigned to a group was 7.

#### *4.2.2. Cases 3 and 4*

In cases 3 and 4, speaker clustering was carried out by using *k-means* with L2 metric, and the solution was assessed by using a deflection coefficient, with the provision that the number of speakers in each group must be not smaller than 6. Such a limitation results from a concern for the quality of the acoustic models. As a consequence, the numbers of speakers in particular groups in the best solution were 9, 9, and 6. Then four sets of acoustic models were created like in cases 1 and 2 and the FER was determined. For case 3, the number of speakers assigned to each group was 4, and for case 4 it was 7. The choice of speakers assigned to particular groups was based on the distance from the mean value. The results of frame recognition effectiveness are shown in columns 6 (case 3) and 7 (case 4) of Table 2.

#### *4.2.3. Conclusions*

The results shown in Table 2 allow formulating the following conclusions:

- When using one set of models for frame content recognition, the difference in recognition quality (FER) for the training and test sets, for all vowels and for the phoneme *a*, is about 5%. Unexpectedly, this difference is the biggest for the vowel *o*. Although the presented measures are averaged it is justifiable to apply the commonly known conclusion that the participation of a given speaker in the training set means better recognition quality for this speaker.
- When using 3 model sets in the training set, a marked improvement in the quality of frame recognition is observed. This improvement, for all the vowels jointly, for the vowel *a*, but also for all the phonemes is about 5%. It is significant. Interestingly, it is the strongest not for *a*, but for *u*. It is noteworthy that this improvement concerns absolutely all vowels, although clustering was performed only based on the phoneme *a*. This points to a strong correlation in vowel articulation, probably related to the size of the vocal tract.
- For the test set, for clustering case 1 using 4 model sets, recognition is better for the phoneme *a*, but also for the phoneme *o*, which has a similar manner of articulation. At the same time, one can observe a deterioration in recognition quality for the remaining vowels, although vowels taken together are slightly better recognised. All phonemes taken together are recognised worse though.
- Clustering cases 3–4 result in a slightly lower quality of frame content recognition. Thus, the proposed UBM-based clustering method displays better properties than the employed *k-means* cases.

### *4.3. Variant 2 – clustering based on a large a observation set*

Columns 6–8 of Table 1 specify UBM weight adaptation results for the vowel *a*, like in experiment 1 with the difference that the adaptation was based on all the observations of *a* from the training set. In such a case, the number of observations is more than 800 for every speaker, and observations depend not only on the speaker but also on the phonetic context. This study has a comparative character and it aims at finding out if about a dozen observations of the word *tak* are representative enough of such a large observation set. There are significant differences in the values of weighting factors in columns 3–5 and 6–8, but for most speakers, the position of the maximum values is identical. The differences in speaker assignment to groups concern 2 speakers from the training set and 3 speakers from the test set. These are small differences then, and the agreement of speaker membership in groups in a set of 36 speakers is 86%. As a result, after dividing speakers into groups, the number of speakers in group 1 is 10, in group 2 is 6 (and the composition of this group is identical to cases 1 and 2) and in group 3 is 8. When checking the effectiveness of frame content

recognition for the test set, the thresholds of group assignment were left unchanged.

Table 3 presents FER indices of recognition quality, which are a consequence of such clustering. A comparison of the results from Tables 2 and 3, for UBM method, reveals slightly better recognition results for the phoneme *a* in the case of clustering based on a large observation set, but these are not large differences. The same remark refers to all the vowels taken together and all the phonemes jointly. At the same time, deterioration in the recognition quality for the phonemes *u* and *o* (back position of the tongue) was observed. Like in experiment 1, one can say that recognition results with the use of the *k-means* method are slightly worse than those for the UBM method.

Table 3. Values of FER as a recognition quality measure for the case without speaker clustering and for different clustering variants, using observations of the phoneme *a* from all the words in the training set.

| phonemes | set | case 0 | case 1 | case 2 | case 3 | case 4 |
|---|---|---|---|---|---|---|
| i | training | 26.1% | 19.8% | | 19.1% | |
| i | test set | 26.0% | 25.9% | 25.4% | 24.6% | 25.0% |
| ɨ | training | 37.3% | 28.3% | | 27.7% | |
| ɨ | test set | 42.4% | 41.6% | 42.5% | 44.0% | 44.1% |
| e | training | 34.7% | 29.7% | | 30.0% | |
| e | test set | 41.1% | 39.8% | 38.5% | 41.8% | 42.4% |
| a | training | 20.2% | 15.8% | | 16.5% | |
| a | test set | 25.2% | 24.6% | 25.2% | 27.3% | 28.1% |
| o | training | 26.4% | 21.8% | | 21.1% | |
| o | test set | 32.7% | 33.0% | 34.6% | 33.8% | 34.4% |
| u | training | 41.6% | 33.6% | | 31.4% | |
| u | test set | 44.9% | 45.5% | 44.8% | 44.8% | 43.8% |
| vowels | training | 28.7% | 23.2% | | 23.1% | |
| vowels | test set | 33.6% | 33.1% | 33.3% | 34.7% | 35.1% |
| all | training | 33.1% | 27.6% | | 27.6% | |
| all | test set | 40.2% | 40.6% | 40.9% | 41.4% | 41.8% |

### *4.4. Variant 3 – clustering based on all vowels*

In general, after applying UBM technique for clustering operation with $F$ phonemes we have available after adaptation process $F \cdot P \cdot G$ weighting coefficients $\mathbf{W} = \{w_{fpg} : f = 1, ..., F; p = 1, ..., P; g = 1, ..., G\}$ where $F$ is a number of phonemes, $P$ is the number of speakers and $G$ is a group number. From the point of view of impossibility of direct utilisation of weighting coefficients clustering algorithm should be extended with additional elements. In fact, executing adaptation process for all phonemes independently, it is almost sure that maximum values of coefficients for different phonemes of the same speaker will occur for different groups, e.g., for a speaker $p$ and phoneme 1 maximum value will have coefficient $w_{1p1}$ but for phoneme 2 coefficient $w_{2p3}$. Random mixing of basis distributions in UBM model is introduced with EM algorithm, however, as is shown in Subsec. 3.1, the maximum value of coefficients implicate membership of a given group. Obtainment of effective clustering procedure forces rearrangement of column vectors in all matrices of the coefficients $\{\mathbf{W_f} : f = 1, ..., F\}$ . The mentioned change of columns' sequence should provide a maximum correspondence of coefficients for each speaker in all groups. The problem is trivial for one spekaer but more complicated in the case of a large number of speakers. For one ilustrative purpose, in Table 4 exemplary coefficient values for 3 phonemes and 7 speakers are presented.

Table 4. Coefficient values of GMM models of phonemes *e*, *a*, *o* after adaptation for 7 speakers.

| speaker | $\mathbf{W_3}$ – phone e | | | $\mathbf{W_4}$ – phone a | | | $\mathbf{W_5}$ – phone o | | |
|---|---|---|---|---|---|---|---|---|---|
| | $w_{3p1}$ | $w_{3p2}$ | $w_{3p3}$ | $w_{4p1}$ | $w_{4p2}$ | $w_{4p3}$ | $w_{5p1}$ | $w_{5p2}$ | $w_{5p3}$ |
| mbi05 | 0.43 | 0.21 | 0.35 | 0.00 | 0.79 | 0.21 | 0.46 | 0.50 | 0.04 |
| mka03 | 0.20 | 0.80 | 0.00 | 0.83 | 0.16 | 0.01 | 0.40 | 0.00 | 0.60 |
| mle01 | 0.32 | 0.68 | 0.00 | 0.89 | 0.11 | 0.00 | 0.21 | 0.00 | 0.79 |
| mle06 | 0.16 | 0.18 | 0.66 | 0.06 | 0.85 | 0.09 | 0.86 | 0.00 | 0.14 |
| mle09 | 0.34 | 0.66 | 0.00 | 0.30 | 0.70 | 0.00 | 0.56 | 0.03 | 0.42 |
| mlu08 | 0.13 | 0.87 | 0.00 | 0.90 | 0.09 | 0.01 | 0.63 | 0.00 | 0.37 |
| mol03 | 0.56 | 0.38 | 0.06 | 0.14 | 0.86 | 0.00 | 0.57 | 0.13 | 0.30 |

In the procedure of column vectors rearrangement we have used crosscorrelation coefficients (inner product) defined as

$$\rho_{i,j,k,l} = \mathbf{w}_{ik}^{\mathrm{T}} \mathbf{w}_{jl}, \tag{17}$$

where $i$ and $k$ are phoneme and cluster numbers of the first element vector and $j$ and $l$ of the second element, respectively. Calculating $\rho_{1,2,2,1}$ coefficient gives a possibility to find evaluation solution of phoneme 1 with vector $\mathbf{w}_{12}$ and phoneme 2 with vector $\mathbf{w}_{21}$.

In view of the problems is complexity, our further considerations concern constructing a global cost function for evaluation of different sequences of column vectors and are limited to the case of $G = 3$ groups. In the first step we consider a choice in matrices $\mathbf{W_f}$ of different variants (variations) of column sequences given in the following form:

$$\boldsymbol{\alpha} = (\alpha(1), \alpha(2), ..., \alpha(F)), \tag{18}$$

where $\alpha(f)$ is a column index chosen in matrix $\mathbf{W_f}$. The total number of different possible variations is $N_1 = G^F = 3^F$ and exemplary arrangement for $F$=6 and $G$=3 can be written as $\boldsymbol{\alpha} = (2, 1, 3, 1, 2, 3)$. For each sequence $\boldsymbol{\alpha}$ we calculate the cost function

$$\overline{\rho}_1(\boldsymbol{\alpha}) = \sum_{i=1}^{F} \sum_{j=i+1}^{F} \rho_{i,j,\alpha(i),\alpha(j)}. \tag{19}$$

It requires a computation of sum of $\binom{F}{2}$ inner products defined by (17). In the second step, from the remaining

columns of matrices $\mathbf{W_f}$, among $N_2 = (G-1)^F = 2^F$ possibilities, we consider the next column sequence $\boldsymbol{\beta} = (\beta(1), \beta(2), ..., \beta(F))$ with the cost function

$$\overline{\rho}_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{F} \sum_{j=i+1}^{F} \rho_{i,j,\beta(i),\beta(j)}. \tag{20}$$

More precisely, if vector $\mathbf{w}_{12}$ was included in the sequence $\boldsymbol{\alpha}$ then it is not taken into consideration in the next step that is in the sequence $\boldsymbol{\beta}$. In the last step we have only one column sequence $\boldsymbol{\gamma} = (\gamma(1), \gamma(2), ..., \gamma(F))$ with the associated cost function

$$\overline{\rho}_3(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{F} \sum_{j=i+1}^{F} \rho_{i,j,\gamma(i),\gamma(j)}. \tag{21}$$

Next we introduce a global cost function related to the joint choice of column sequences $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$

$$\overline{\rho}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \overline{\rho}_1(\boldsymbol{\alpha}) + \overline{\rho}_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \overline{\rho}_3(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \tag{22}$$

and, among $N = 6^F$ different possibilities, we search for the maximum rearrangement

$$(\boldsymbol{\alpha}_{opt}, \boldsymbol{\beta}_{opt}, \boldsymbol{\gamma}_{opt}) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}} \overline{\rho}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}). \tag{23}$$

The obtained form of the optimal solution $(\boldsymbol{\alpha}_{opt}, \boldsymbol{\beta}_{opt}, \boldsymbol{\gamma}_{opt})$ differs from the initial configuration with ordered groups for individual phonemes according to the considered cost function (23). Finally, for a given optimal vectors rearrangement $\mathbf{w}_{fg}$, denoted as $\widetilde{\mathbf{w}}_{fg}$, the averaging operation with respect to phonemes is introduced for an individual speaker with index $p$

$$\overline{w}_{pg} = \frac{1}{F} \sum_{f=1}^{F} \widetilde{w}_{fpg}. \tag{24}$$

As a result, we obtain an individual decision matrix $[\overline{\mathbf{w}}_{\mathbf{p1}}\ \overline{\mathbf{w}}_{\mathbf{p2}}\ \overline{\mathbf{w}}_{\mathbf{p3}}]$, where classification formula of the maximum type can be executed. Coefficients $\overline{w}_{pg}$ represent all the considered phonemes. A generic scheme representing the choice of the best rearrangement of column vectors of coefficients for $F$ phonemes and procedure of computation of the averaged coefficients $\overline{w}_{pg}$ for $G = 3$ is depicted in Fig. 4.

Columns 9–11 of Table 1 present the results of the averaged weighting factors determined in this way. Like in experiments 1 and 2, these factors form the basis for division of the speakers in the training set into groups and assigning the test set speakers to groups. As Table 1 suggests, the agreement of speaker membership in groups in relation to experiment 1 is about 64%, which, allowing for the use of a slightly different methodology and 3 groups, indicates a good agreement. While dividing the speakers in the training set into groups, it turned out that the size of

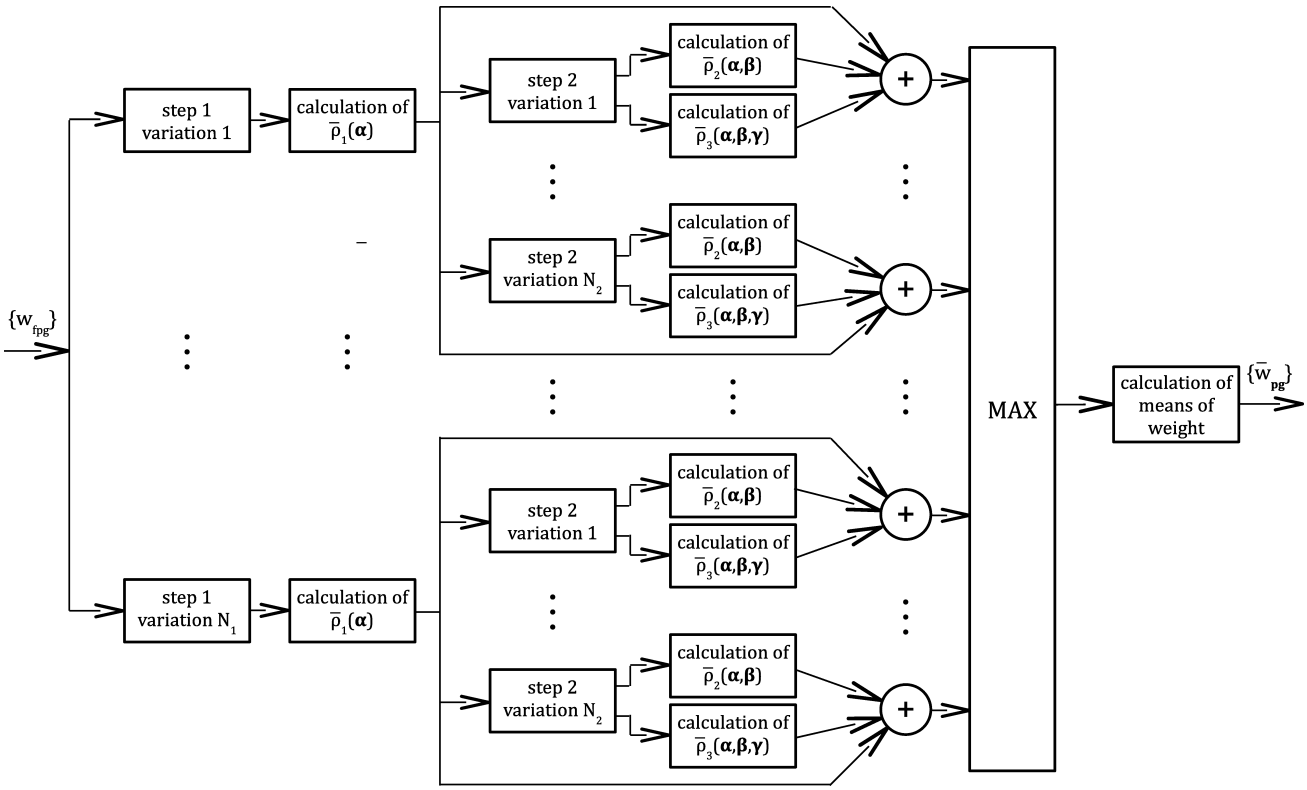


Fig. 4. Illustration of the choice of the optimal rearragement of coefficient vectors for phonemes to groups and computation of the averaged coefficients $\overline{w}_{pg}$ for $G = 3$.

group 2 is too small (< 6). Therefore, two speakers from group 3 (speakers mle06 and mry09) were transferred to this group. As a result, the number of speakers in group 1 became 11, in group 2 it was 6, and in group 3 it was 7.

Table 5 presents the values of FER used as a recognition quality measure resulting from UBM-based clustering. A comparison of the results from Tables 2 and 5 (experiments 1 and 3) indicates that when it comes to all the vowels taken together and all the phonemes together, differences in recognition quality are negligible. At the same time, the results for the vowel *a* in the test set are significantly worse, while these obtained for the remaining vowels are better and they meet the expectations. Generally, the conclusions from experiment 3 are comparable to those from experiment 1.

Table 5. Values of FER as a recognition quality measure for the case without speaker clustering and for UBM clustering, when using observations of all the vowels from all the words in the training set.

| phonemes | set | case 0 | case 1 | case 2 |
|---|---|---|---|---|
| i | training | 26.1% | 20.0% | |
| i | test set | 26.0% | 25.8% | 25.4% |
| ~~i~~ | training | 37.3% | 29.0% | |
| ~~i~~ | test set | 42.4% | 44.5% | 42.9% |
| e | training | 34.7% | 29.8% | |
| e | test set | 41.1% | 38.6% | 38.0% |
| a | training | 20.2% | 16.2% | |
| a | test set | 25.2% | 27.5% | 28.5% |
| o | training | 26.4% | 21.3% | |
| o | test set | 32.7% | 31.1% | 32.6% |
| u | training | 41.6% | 33.6% | |
| u | test set | 44.9% | 44.2% | 45.3% |
| vowels | training | 28.7% | 23.4% | |
| vowels | test set | 33.6% | 33.5% | 33.8% |
| all | training | 33.1% | 27.8% | |
| all | test set | 40.2% | 40.8% | 41.1% |

## 5. Summary

The report has presented a new speaker clustering method based on the universal background model, which is known from the speaker identification problem. When using this model for clustering, it was proposed that adaptation of UBM model parameters to a given utterance will be used only for the weighting factors of the acoustic GMM model. The proposed adaptation method is very fast, which enables it to be used not only at the ASR systems training stage (clustering) but also at recognition stage with the aim of assigning a given utterance (speaker) to one of several acoustic models.

The application of clustering for the cases represented in the conducted experiments results in the improvement of the training set frame content recognition quality for all vowels jointly and for all phonemes jointly by over 5%. As for the test set, this improvement is smaller, which results from (one could realise that when analysing adapted weight values) underrepresentation of particular speakers in UBM models. This, in turn, may be due to a too small speaker set, which results in the insufficient number of groups and sets in GMM models. The proposed clustering method is more effective than the classic *k-means* method with different metrics.

A very interesting conclusion from the conducted research is that clustering based on the phoneme *a* results in an improvement in the recognition quality for all vowels. This proves a similarity in the mode of articulation of all vowels by a particular speaker, which is probably related to the size of the vocal tract. A similar conclusion about recognition effectiveness is true for all phonemes jointly, but one must be careful with generalising the ensuing conclusions, as the mode of articulation of many consonants (e.g., plosives) is unrelated to the size of the vocal tract.

What is also interesting, is that clustering based on the phoneme *a* from one utterance is as effective as that based on many fragments containing *a* from many different words, as well as clustering based on all vowels. Thus, choosing one of the acoustic model sets based exclusively on the word *tak* is a good solution of improving the effectiveness of all the ASR system. The choice of *a* as the phoneme on which assignment to a group is based is not accidental, as it is the most recognisable vowel in the Polish language.

## Acknowledgment

The research was financed partly by a grant from The National Centre for Research and Development, grant no ID: 245755 (Audioscope).

## References

1. Anderson T.W. (2003), *An Introduction to Multivariate Statistical Analysis*, 3rd ed., John Wiley & Sons Inc, New York.
2. Basseville M. (1989), *Distance measures for signal processing and pattern recognition*, Signal Processing, **18**, 349–369.
3. Bishop C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York.
4. Chu S.M., Tang H., Huand T.S. (2009a), *Locality preserving speaker clustering*, Proceedings of IEEE International Conference on Multimedia and Expo, pp. 494–497, Mexico.
5. Chu S.M., Tang H., Huang T.S. (2009b), *Fisher-voice and semi-supervised speaker clustering*, International Conference on Acoustics, Speech and Signal Processing, pp. 4089–4092, Taipei.

6. De La Torre A., Peinada A.M., Segura J.C., Perez-Cordoba J.L., Benitez M.C., Rubio A.J. (2005), *Histogram equalization of speech representation for robust speech recognition*, IEEE Transaction on Speech and Audio Processing, **13**, 355–366.
7. Duda R., Hart P., Stork D. (2000), *Pattern Classification*, 2-nd ed., John Wiley & Sons Inc., New York.
8. Hazen T.J. (2000), *A comparison of novel techniques for rapid speaker adaptation*, Speech Communication, **31**, 15–33.
9. He X., Niyogi P. (2003), *Locality Preserving Projections*, Advances in Neural Information Processing Systems, **16**, Vancuver.
10. Iyer A.N., Ofoegbu U.O., Yantorno R.E., Smolinski B.Y. (2006), *Blind Speaker Clustering*, International Symposium on Intelligent Signal Processing and Communications Systems, pp. 343–346, Yonago.
11. Jassem W. (1973), *Fundamentals of Acoustic Phonetics* [in Polish: *Podstawy fonetyki akustycznej*], PWN, Warszawa.
12. Kosaka T., Sagayama S. (1994), *Tree-structured speaker clustering for fast speaker adaptation*, Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 245—248, Ostendorf.
13. Kuhn R., Junqua J.-C., Nguyen P., Niedzielski N. (2000), *Rapid speaker adaptation in eigenvoice space*, IEEE Transaction on Speech and Audio Processing, **8**, 695–707.
14. Liu D., Kubala F. (2004), *Online Speaker Clustering*, Procedings of International Conference on Acoustics, Speech and Signal Processing, pp. 333–336, Quebec.
15. Lu Z., Hui Y.V., Lee A.H. (2003), *Minimum Hellinger distance estimation for finite Poisson regression models and its applications*, Biometrics, **59**, 1016–1026.
16. Mehrabani M., Hansen J.H.L. (2013), *Singing speaker clustering based on subspace learning in the GMM mean supervector space*, Speech Communication, **55**, 653–666.
17. Makowski R. (2011), *Automatic speech recognition – selected problems* [in Polish: *Automatyczne rozpoznawanie mowy – wybrane zagadnienia*], Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.
18. Makowski R., Hossa R. (2014), *Automatic speech signal segmentation based on innovations adaptive filter*, International Journal on Applied Mathematics and Computer Science, **24**, 259–270.
19. Mrówka P., Makowski R. (2008), *Normalization of speaker individual characteristics and compensation of linear transmission distortions in command recognition systems*, Archives of Acoustics, **33**, 221–242.
20. Naito M., Deng L., Sagisaka Y. (2002), *Speaker clustering for speech recognition using vocal track parameters*, Speech Communication, **36**, 305–315.
21. Reynolds D.A., Rose R.C. (1995), *Robust text-independent speaker identification using gaussian mixture speaker models*, IEEE Transaction on Speech and Audio Processing, **3**, 72–83.
22. Reynolds D.A., Quatieri T.F., Dunn R.B. (2000), *Speaker verification using adaptive gaussian mixture models*, Digital Signal Processing, **10**, 19–41.
23. Stafylakis T., Katsouros V., Carayannis G. (2006), *The segmental Bayesian Information Criterion and its applications to Speaker diarization*, IEEE Selected Topics in Signal Processing, **4**, 857–866.
24. Tang H., Chu S.M., Hasegawa-Johnson M., Huang T.S. (2012), *Partially Supervised Speaker Clustering*, IEEE Transaction on Pattern Analysis and Machine Intelligence, **34**, 959–971.
25. Tranter S., Reynolds D. (2006), *An overwiew of Autmatic Speaker Diarization Systems*, IEEE Transaction Audio, Speech and Language Processing, **14**, 1557–1565.
26. Tsai W-H., Cheng S-S., Wang H-M. (2007), *Automatic Speaker Clustering Using a Voice Characteristic Reference Space and Maximum Purity Estimation*, IEEE Transaction on Audio, Speech and Language Processing, **15**, 1461–1474.