ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20140120095957

SPEAKER MODEL ADAPTATION BASED ON CONFIDENCE SCORE

Erhan Mengusoglu

Original scientific paper
Confidence measures are expected to give a measure of reliability on the result of a speech/speaker recognition system. Most commonly used confidence measures are based on posterior word or phoneme probabilities which can be obtained from the output of the recognizer. In this paper we introduced a linear interpretation of posterior probability based confidence measure by using inverse Fisher transformation. Speaker adaptation consists in updating model parameters of a speaker independent model to have a better representation of the current speaker. Confidence measures give more reliable selection criteria to select the utterances which best represent the speaker. A linear interpretation of confidence measure is very important to select the most representative data for adaptation.

Keywords: *confidence measure; Fisher transformation; speaker adaptation; speaker verification*

Adaptacija modela govornika na osnovu rezultata povjerenja

Izvorni znanstveni članak
Očekuje se da mjere povjerenja postanu mjera za pouzdanost rezultata sustava za prepoznavanje govora. Najčešće korištene mjere povjerenja zasnovane su na vjerojatnosti sljedeće riječi ili fonema, koja se može dobiti iz izlaznog rezultata prepoznavatelja. U ovom smo radu uveli mjeru povjerenja zasnovanu na linearnoj interpretaciji vjerojatnoće sljedeće riječi primjenom obrnute Fisher transformacije. Adaptacija govornika sastoji se od ažuriranja parametara modela nezavisnog od govornika zbog boljeg predstavljanja postojećeg govornika. Mjere povjerenja daju pouzdanije kriterije za odabir riječi koje najbolje predstavljaju govornika. Linearna interpretacija mjere povjerenja vrlo je važna pri odabiru najreprezentativnijih podataka za adaptaciju.

Ključne riječi: *adaptacija govornika; Fisherova transformacija; mjera povjerenja; verifikacija govornika*

1 Introduction

Improving the accuracy of a speech/speaker recognition system is a major research area in the current speech technology researches. Current speech/speaker recognition systems are not accurate enough when they are used in conditions different from the training conditions of the systems. The main research directions to improve the accuracy of systems are:

- Noise robust recognition
- Improved feature extraction
- Confidence measures
- Speaker model adaptation
- Improved language modelling

This paper is interested in the use of confidence measures in speaker model adaptation for the speaker verification task. Confidence measures are used on various applications in speech recognition field. Experiments in large vocabulary continuous speech recognition, reported in [1] and [2], show that the use of confidence measure for constructing a word graph significantly increases the recognition performance. In [4] application of confidence measure in language identification task is explained.

Confidence measures are generally based on posterior probabilities of the recognition system. Use of some prior information [5] has shown that it improves the efficiency of confidence measures. Some confidence measures are based on language model probabilities [6]. In this case a word graph is constructed from language model probabilities and the word sequence obtained by the recognizer is scored by this graph.

As discussed in the survey provided in [15], majority of confidence measures rely on posterior probability computed using forward-backward algorithm. The confidence on the recognized utterance will be high if the utterance in the recognized path is significantly different from the competing graphs. If the difference between recognized path and the competing paths is not important then the confidence score will be low.

2 Speaker model adaptation for speaker verification

Speaker verification is a biometric technique used for improving security in access control. The basic assumption behind the use of speaker verification is that, like finger prints, every human has unique speech characteristics that can distinguish a speaker from another. Recently, banks, mobile phone operators and other service providers that propose phone based customer services started to get interested in adding speaker verification for identity verification purposes.

Description of how speaker adaptation is carried out in order to adapt speaker models to changes in the verification environment is provided in this section. We will also provide different techniques that can be applied to adaptation of models. Changes in the verification environment that requires adaption include intra-speaker variabilities (speaking rate, changes related on sickness …) and extra-speaker variabilities (noise, echo …). When there is a mismatch between training conditions of speaker models, the verification system is less accurate.

The speaker verification technique used in this paper is based on Gaussian Mixture Model (GMM) [7] and is independent of phonetic content in the speech.

The structure of speaker verification system used is given in Figs. 1 and 2. There are two types of model; speaker model and world model. Speaker model is trained on few sentences uttered by a speaker and world model is trained by a relatively large data set obtained from different speakers.

In the experiments, one state and two state GMMs are used. There are two types of two state GMMs:

1) One state for silence and one state for speech
2) One state for unvoiced phonemes and one state for voiced phonemes.

Initial speech/silence labelling and voiced/unvoiced labelling is obtained by applying HMM/MLP speech recognition [8] on the world model training data to obtain phoneme probabilities for each frame. These probabilities are then used to label the data. After having obtained a two state world model, it is used to label speaker data then the labelled speaker data is used to train speaker models.

In the verification system, accept/reject decision is taken by comparing the likelihood score [7] of the utterance given the model of claimed speaker with the likelihood score computed for an impostor model. There are two impostor models included in the system, one for female and one for male speakers.

There are two speaker model adaptation techniques used in this paper; Maximum A posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR).

3 Maximum-A-Posteriori (MAP)

Model adaptation using MAP involves prior knowledge about the parameter distribution of the model to be adapted. This prior knowledge is used as a base for the new model which should better model the newly observed data. There is a weighing factor based on the availability of adaptation data and the degree of mismatch between newly observed data and data used to train the initial model. The use of prior information prevents over fitting the original model to the observed data. When the amount of observed data is small, over fitting can cause degradation in accuracy of the system.

In the MAP adaptation, the main goal is to maximize an *aposteriori* function based on likelihoods and prior probabilities. The model parameters are updated to achieve this goal [9]:

$$\lambda_{MAP} = arg\max_{\lambda} f(\lambda|O). \quad (1)$$

By applying Bayes' theorem, this formula takes the following form.

$$\lambda_{MAP} = arg\max_{\lambda} \frac{L(O|P_0(\lambda))}{P(O)}. \quad (2)$$

In this formula O is the observation vector. $L(O|\lambda)$is the likelihood of the observed data given the present model. $P(O)$, the a priori probability of the observed data, is omitted because it does not depend on the model. $P_0$is the prior probability density function of the model. What is tried to be achieved by applying this formula is find a set of parameters that best represent the observed data.

For simplicity, MAP adaptation can be used to adapt only the means of Gaussians in a GMM. The update formula [10] to obtain adapted means for some observation data is defined as follows,

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\overline{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm}, \quad (3)$$

where τ is the weighing factor for prior knowledge, N is the occupation likelihood of the adaptation data defined as,

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t), \quad (4)$$

where R is the number of states, T is the number of observation vectors, μ_{jm} is the mean parameter of the model to be adapted and $\overline{\mu}_{jm}$ is the mean of the observation data defined as,

$$\mu_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)}. \quad (5)$$

As can be seen from the update formula, when the likelihood of observation data is higher, the adaptation level will also be higher. MAP adaptation performs better when more adaptation data is available because more data will help creating a model that represents better the observed data.

4 Maximum Likelihood Linear Regression (MLLR)

The use of MLLR for model adaptation consists in producing a set of regression based transforms from some adaptation data. These transforms are then used to tune the parameters of the GMM to be adapted. MLLR transformations are generally only applied to means of Gaussian which are the most important components of GMMs to be updated when they are tried to be adapted in order to represent new conditions [11].

The use of MLLR for mean transformation of a Gaussian mixture model consists in computing a transformation matrix from observations and then using it to obtain adapted means.

For observations of dimension *n*,

$$\mu_s = W_s\xi_s, \quad (6)$$

where W_s is a transformation matrix of size $n \times (n+1)$ and $\xi_s = [w, \mu_{s1}, \dots, \mu_{sn}]^t$ is the extended mean vector in which w=1 indicates that there is an offset and $w = 0$ means no offset.

$W_s$is computed by solving the following equation;

$$\sum_{t=1}^{T}\sum_{r=1}^{R} L_{sr}(t)\sum_{sr}^{-1} O(t)\xi_{sr}^{T} = \sum_{t=1}^{T}\sum_{r=1}^{R} L_{sr}(t)\sum_{sr}^{-1} O(t) W_s\xi_{sr}\xi_{sr}^{T}. \quad (7)$$

$L_{sr}(t)$ is the occupation likelihood which is obtained from forward backward process. Implementation issues could be found in [9].

5 Use of confidence measure for unsupervised adaptation

Confidence measure used in this paper is based on likelihood ratios obtained for tested utterance given the model for a certain speaker. The measure will provide a confidence score for each of the adaptation utterances. Based on certain confidence threshold, speaker adaptation system will decide if the utterance will be used for adapting the speaker model. When computing the

confidence score, not only speaker model is used but also the score for an impostor model is also computed. The threshold confidence measure is tested against the difference between the two scores obtained from use of speaker model and impostor model. That means likelihood score from speaker model is divided (subtracted in log domain) by likelihood score from impostor model.

$$\Lambda(X) = \log p(X|\lambda_c) - \log p(X|\lambda_{\bar{c}}). \quad (8)$$

$p(X|\lambda_c)$ is the likelihood that utterance X belongs to the claimed speaker and $p(X|\lambda_{\bar{c}})$ is the likelihood that utterance does not belong to the claimed speaker. λ_c is the speaker model and $\lambda_{\bar{c}}$ is the world (background) model.

$\Lambda(X)$ is then compared with claimed speaker Gaussian mean and impostor Gaussian mean. Those Gaussians are computed using the claimed speaker data and impostor data. If the value is between two means then the confidence measure is computed by; first, normalizing the likelihood value by the two Gaussians (10) and (11), then, the normalized values are transformed to correlation domain by using inverse Fisher transformation [12]. This transformation is generally used for determining a confidence interval for the correlation of different data set.

$$z = \frac{1}{2}\log(\frac{1+r}{1-r}). \quad (9)$$

In this formula, z is the transformed value of correlation value r. z has a Gaussian (normal) distribution. Correlation value can give the importance of relation between two data sets.

The values obtained after transformation are the measures of relationship between likelihood ratio and claimed speaker and impostor.

$$z_{\text{speaker}} = \frac{\Lambda(x)-\mu_{\text{speaker}}}{\sigma_{\text{speaker}}} \quad (10)$$

$$z_{\text{impostor}} = \frac{\Lambda(x)-\mu_{\text{impostor}}}{\sigma_{\text{impostor}}} \quad (11)$$

$$r = \frac{e^{2z}-1}{e^{2z}+1} \quad (12)$$

$$cm = r_{\text{speaker}} - r_{\text{impostor}}.$$

cm is the final confidence score that can be used directly for determining which utterance is eligible for speaker model adaptation. If the confidence score is negative, that means the utterance comes from an impostor. If the score is positive, then utterance seems to be pronounced by the true speaker. In this case confidence score determines the level of confidence. For adaptation it is better to use only the utterances with high confidence scores.

6 Experimental setup

The POLYCOST [13] speaker verification database is used for experiments. The database is specifically recorded for speaker verification tasks over the telephone. Speakers from different countries across Europe are involved in recordings, speakers are asked to utter a pre-determined utterance in their mother language. Since the recordings are done over the telephone line, microphone characteristics and SNR values for the database are not available but for the purpose of experiments reported in this paper we consider non-existence of these properties for the database as non-relevant. Database is considered as low SNR even if there are no formal SNR specifications based on listening tests. The results provided show relative improvement in speaker verification tasks when using adaptation.

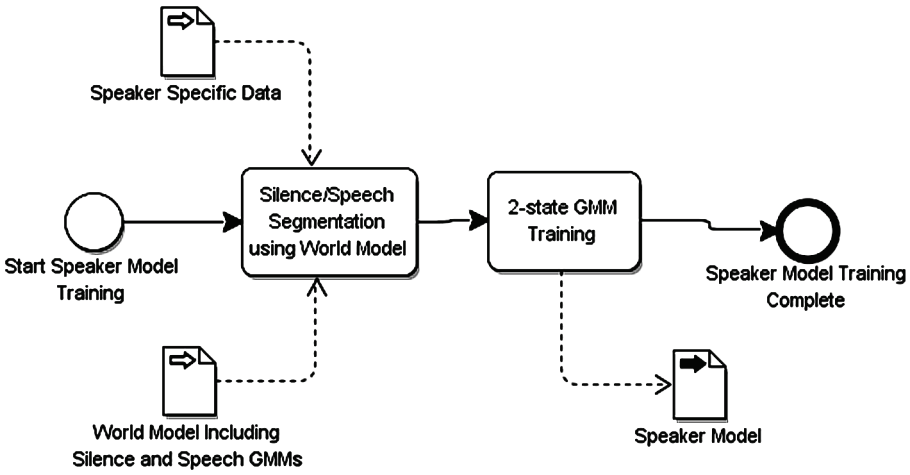

Figure 1 Training process for speaker model

We have created, three sets of data:

- Training,
- Testing,
- Adaptation for each speaker.

Artificially created white Gaussian noise added (SNR = 15 dB) version of each data set is also used. A large amount of data is selected from the database randomly for training of a *"world model"* and created two other subsets of data for training *"impostor models"* (one for females and one for males). Noisy versions of the data sets are also created for noise tests. Since the results for female speakers and male speakers are different, test results are listed separately for females and males.

Figs. 1 and 2 show the training process and verification processes using BPMN diagrams. GMMs

used in the experiments have 2 states (speech and silence). By separating silence and speech we aimed at basing our score computations only on the speech parts of the utterances which carry the most speaker specific information.

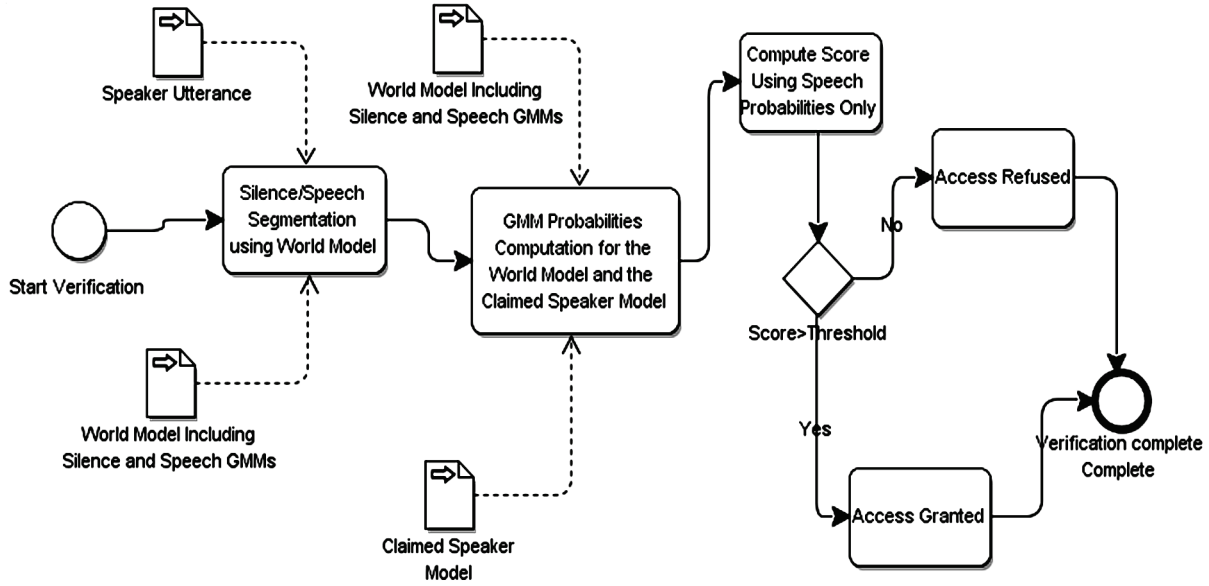

Figure 2 Speaker verification setup

7 Results and discussions

Some results obtained with the techniques explained in previous sections are provided in this section. The results documented in this section show how the use of confidence measures for unsupervised adaptation increases verification performance. Some other results for different experiments are carried out for testing different combinations of speaker segmentation techniques. Note that the results provided here are showing only relevant improvements in verification performances. There are no comparisons with the baseline systems reported elsewhere.

Three different types of experiment are realized to test the effect of adaptation on the accuracy improvement of speaker verification system.

1. Use of 1 state GMM for each speaker and the world model.
2. Use of 2 states GMM, one for speech and one for silence
3. Use of 2 states GMM, one for voiced phonemes and one for unvoiced phonemes.

There are six test groups. For every test, there is an explanation followed by a table. The tables are identical in form. There are four columns in the tables:

- Type of test
- Gender
- False rejection error rate
- False acceptance error rate.

In Tab. 1, when training and test data are both clean, best verification results are obtained with voiced/unvoiced modelling for female speakers and speech/silence modelling for male speakers.

In Tab. 1, it is clear that the use of adaptation does not improve the verification results when there is no mismatch between training and test data. In this example both data sets are clean data. This makes sense because trying to apply adaptation to a well-trained model will wrongly move the model towards the adaptation data. This is why adapting a well-trained model with new data without having confidence on whether the new data is appropriate for adaptation is a bad idea. The results here confirm the value of our work in this paper. We clearly see that selecting the utterance to be used for adaptation is important. In this first task, an adaptation mechanism that does not take into account the quality of adaptation utterance could result in worse performances.

Table 1 Clean data, use of 1 state GMM (res), 2 state GMM for silence-speech (res-sil) and two state GMM for unvoiced/voiced (res-uv)

Type of test	Male/female	Fault rejection error rate / %	Fault acceptance error rate / %
res	female	3,46	2,95
	male	4,69	4,12
res-sil	female	3,21	2,70
	male	3,70	3,04
res-uv	female	3,09	3,02
	male	4,07	3,54

Table 2 MLLR adaptation and MAP adaptation with clean adaptation data

Type of test	Male/female	Fault rejection error rate / %	Fault acceptance error rate / %
res_mllr	female	6,91	5,44
	male	5,06	3,95
res_map	female	10,74	8,34
	male	10,25	8,22
res_mllr-sil	female	8,15	6,45
	male	4,44	3,43
res_mllr-uv	female	7,04	6,08
	male	4,44	3,68

Tab. 3 shows that, when there is a mismatch between training and test conditions (noisy test data), silence/speech labelling performs better for male speakers and 1-state modelling performs better for female speakers.

This table would be our new benchmark from the following tests as adaptation makes sense when there is a mismatch between testing and training conditions.

Table 3 Use of noisy test data with GMMs trained on clean data. Noise (SNR=15) was added to the test data used on the previous tables

Type of test	Male/female	Fault rejection error rate / %	Fault acceptance error rate / %
res-n15	female	13,33	9,82
	male	9,14	9,50
res-sil-n15	female	16,91	1,32
	male	8,64	8,50
res-uv-n15	female	14,44	1,15
	male	10,00	9,84

Table 4 Adaptation of GMMs trained on clean data with noisy adaptation data. MAP adaptation is only applied on single GMM case

Type of test	Male/female	Fault rejection error rate / %	Fault acceptance error rate.
res_mllr-n15	female	13,58	10,16
	male	5,19	3,73
res_map-n15	female	14,57	8,52
	male	10,49	8,59
res_mllr-sil-n15	female	13,33	12,79
	male	5,93	4,39
res_mllr-uv-n15	female	12,47	10,51
	male	6,91	4,54

Tab. 4 shows that MLLR adaptation of single state GMMs works better than adaptation of two state GMMs and also use of MAP adaption is not generating good results compared with MLLR adaptation. Best performing adapted models are the male speaker models adapted using MLLR. The reason for MAP performing worse could be insufficient amount of data used for adaptation. MAP is known to require a large amount of data as it is purely based on estimating statistical distribution of the data where MLLR is simply computing regression parameters.

Considering better performances with MLLR adaptation, confidence measure based adaptation for the rest of the experiments will only be applied to MLLR adaptation. Poor results for MAP adaptation might be because of MAP method needing more adaptation data.

Table 5 Use of confidence measure (confidence threshold=0,5) to select the adaptation utterance

Type of test	Male/female	Fault rejection error rate / %	Fault acceptance error rate / %
res_mllr-n15-cm	female	12,10	8,11
	male	5,56	4,53
res_mllr-sil-n15-cm	female	12,96	11,44
	male	6,91	5,84
res_mllr-uv-n15-cm	female	11,73	9,23
	male	7,78	6,07

Tab. 5 shows that use of confidence measure driven adaptation improves the verification performance of female speaker models. The results reported on this table should be compared to the results in Tab. 6 where noisy training data is used in the first place.

In Tab. 6, since the training and test conditions match, ideally, the best verification performance for noisy test data would be obtained. Comparing Tab. 5 and Tab. 6 we can see that the adapted female speaker model performs similar to the model trained with noisy data and the adapted male speaker model performs better than the model trained with noisy data. When compared with the performance of female models in Tab. 4, we can clearly see that use of confidence measure helps selecting the data that would lead to a better adaptation.

Table 6 Use of noisy training data

Type of test	Male/female	Fault rejection error rate / %	Fault acceptance error rate / %
res-n15	female	11,85	7,04
	male	7,16	4,20
res-sil-n15	female	13,82	10,71
	male	9,51	5,93
res-uv-n15	female	13,95	8,99
	male	8,77	5,76

8 Conclusion

The results reported show that confidence measure driven MLLR adaptation improves significantly the speaker verification performances. We have observed that use of two state GMMs in either speech-silence or voiced-unvoiced methods with MLLR adaptation to selected utterances works well. MAP adaptation does not perform better than MLLR adaptation but this may be due to limited number of adaptation data. Combined MAP+MLLR adaptation needs to be investigated. Other interesting investigation areas could be use of the world model as a base for adaptation, use of varying prior knowledge weighing as a function of amount of the available adaptation. The latter could be done by decreasing the weight factor in MAP adaptation, τ in Eq. (3), when there are more utterances to be used for adaptation.

Finally we could confirm that use of confidence measure for selecting adaptation data will prevent the over fitting of the adapted speaker model which will result in performance degradation. Recent research on use of GMM for speaker verification [14] confirms the validity of the experiments carried out in this paper. Further investigation is needed to evaluate the validity of the technique provided here on real-word speaker verification tasks. Speaker verification over the telephone remains a topic of interest in real world applications.

9 References

[1] Wessel, F.; Schlüter, R.; Macherey, K.; Hey, H. Confidence Measures for Large Vocabulary Continuous Speech Recognition. // IEEE Transactions on Speech and Audio Processing. 9, 3(2001), pp. 288-298. DOI: 10.1109/89.906002

[2] Cox, S. High Level Approaches to Confidence Estimation in Speech Recognition. // IEEE Transactions on Speech and Audio Processing. 10, 7(2002), pp. 460-471. DOI: 10.1109/TSA.2002.804304

[3] Skantze, G. The use of Speech Recognition Confidence Scores in Dialogue Systems, Goteborg University, Graduate School of Language Technology, Speech Technology 1, course term paper, 2003.

[4] Metze, F.; Kemp, T.; Schaaf, T.; Schultz, T.; Soltau, H. Confidence Measure Based Language Identification, ICASSP 2000, Istanbul, Turkey, 2000.

[5] Mengusoglu, E.; Ris, C. Use of Acoustic Prior Information for Confidence Measure in ASR Applications, EuroSpeech 2001, Aalborg, Denmark, 2001.

[6] Hacioglu, K.; Ward, W. A Concept Graph Based Confidence Measure, ICASSP 2002, Orlando-Florida, USA, 2002.
[7] Reynolds, D. A. Automatic Speaker Recognition Using Gaussian Mixture Speaker Models. // The Lincoln Laboratory Journal. 8, 2(1995), pp. 173-192.
[8] Bourlard, H.; Morgan, N. Connectionist Speech Recognition: A Hybrid Approach, Kluwer, 1994. DOI: 10.1007/978-1-4615-3210-1
[9] Nguyen, P. Fast Speaker Adaptation, Technical Report, Eurecom, 1998.
[10] Young, S.; Evermann, G.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Valtchev, V.; Woodland, P. The HTK book for HTK Version 3.1 (Cambridge University Engineering Department, December 2001).
[11] Hamaker, J. E. MLLR: A Speaker Adaptation Technique for LVCSR, Lecture for a course at ISIP - Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University, 1999.
[12] Fisher, R. A. Statistical Methods Experimental Design and Scientific Inference (Oxford Science Publications, 1890-1962).
[13] Petrovska, D.; Hennebert, J.; Melin, H.; Genoud, D. Polycost: A Telephone-Speech Database for Speaker Recognition. // Speech Communication. 31, 2-3(2000), pp. 265-270.
[14] Kellya, F.; Drygajlob, A.; Hartea, N. Speaker verification in score-ageing-quality classification space. // Computer Speech & Language. 27, 5(2013), pp. 1068-1084. DOI: 10.1016/j.csl.2012.12.005
[15] Jiang, H. Confidence measures for speech recognition: A survey. // Speech communication. 45, 4(2005), pp. 455-470. DOI: 10.1016/j.specom.2004.12.004

Author's addresses

Erhan Mengusoglu, Assistant Professor
TED University, Computer Engineering Department
Ziya Gokalp Caddesi No 48, Kolej, Cankaya,
Ankara, Turkey
E-mail: mengusoglu@gmail.com