Self-Organizing Feature Map Preprocessed Vocabulary Renewal Algorithm for the Isolated Word Recognition System

A. Serackis[1], G. Tamulevicius[1,2], T. Sledevic[1], L. Stasionis[1], D. Navakauskas[1]
[1]*Department of Electronic Systems, Vilnius Gediminas Technical University, Naugarduko St. 41–413, LT-03227 Vilnius, Lithuania*
[2]*Institute of Mathematics and Informatics, Vilnius University, Akademijos St. 4, LT-08663 Vilnius, Lithuania*
arturas.serackis@vgtu.lt

***Abstract*—Paper focuses on the new vocabulary renewal algorithm designed for the hardware implemented Lithuanian speech recognizer. The isolated word recognition is performed using dynamic time warping of the Mel-frequency cepstrum coefficients (MFCC) estimated during short-time analysis of speech signals. A self-organizing feature map is used to extract the time-dependent MFCC features variations. To increase the isolated word recognition rate, four references are stored in the vocabulary for each word to be recognized. In order to make vocabulary adaptive to long-term changes of the user speech and adapt recognizer to the environment the references should be updated. The renewal of the vocabulary is performed if two conditions are met: the distance between same word references and the distance between new reference and other word references in the feature set should be increased. The comparison of the time-dependent MFCC feature variations is performed using Needleman-Wunsch sequence alignment algorithm.**

***Index Terms*—Home automation, human computer interaction, automatic speech recognition, self-organizing feature maps, field programmable gate arrays.**

I. INTRODUCTION

In this paper the new vocabulary renewal algorithm is proposed. The algorithm is applied to the isolated word recognition system of Lithuanian speech [1]. The system is implemented in hardware using Field Programmable Gate Arrays (FPGA) and is able to perform isolated word recognition in real-time.

Due to variations of background noise and pronunciation of the same word the practical application of the speech recognition system adds additional pattern matching challenges comparing the new incoming word pattern and reference pattern stored in the vocabulary [2]. The dynamic time warping (DTW) based isolated word recognition algorithm becomes sensitive to the time-dependent variations of the features, estimated for the word, if the duration of the new pattern differs 1.5–2 times comparing to the reference stored in the vocabulary, especially if the vocabulary stores similar words with the difference of one or two phonemes [3]. In order to reduce the sensitivity of the recognizer to the differences of the same word pronunciation, the vocabulary is supplemented by several examples of the same word pronunciation – additional references are added to the vocabulary. However it is unclear which additional pronunciations should be stored as additional references in the vocabulary if there are more of them than there are possible to store in the vocabulary [4]. Another challenge is to continuously update the vocabulary during the operation of the speech recognizer in order to follow the long-term changes in word pronunciation by the system user [5].

Many methods in automatic speech recognition and word labelling are supervised and need manual labelling of recognized word. With respect to the time-consuming training there are developed word selection methods like grammar-based semi-supervised incremental learning [6].

For the aging speech recognition the maximum likelihood linear regression and maximum a posteriori of the speech features are used to improve recognition accuracy [7].

The dual memory system was proposed in which new information about spoken word is initially encoded separately from existing knowledge and integrated with long-term vocabulary over time [8].

In the following proposed algorithm two main tasks are solved: simplification of the reference comparison and decision making in order to update the vocabulary. The decision on the reference removal and addition of the new pattern to the vocabulary is made accordingly to the two rules:

R1. The new pattern should be less similar to the same word references already stored in the vocabulary.
R2. The addition of the new pattern should not decrease the distance between vocabulary references in the feature set.

The proposed algorithm makes the speech recognition system adaptive and continuously self-updating system, able to adapt to the word pronunciation changes that appear in the user's speech.

The experimental investigation has shown the increase of

Manuscript received November 28, 2013; accepted February 19, 2014.
This research was funded by a grant (No. MIP-092/2012) from the Research Council of Lithuania.

the recognition rate if the proposed vocabulary renewal algorithm is applied to the system. The previous evaluation of different speech analysis methods shows that MFCC features gives highest recognition accuracy in developed hardware-based speech recognizer [9]. The performance of speech recognition systems with MFCC analysis is considered to be good in noise-free environment [10], [11].

II. Isolated Word Recognition System

The pattern matching performed in speech recognition system is based on and adopted for the implementation in FPGA based real-time system [12]. Speech signal analysis (based on acoustical model) is performed in four steps. The averaging filter is applied to the 44.1 kHz sampled recordings for the averaging of four adjacent samples thus obtaining 11 kHz sampled speech records [12]. This allows reducing the influence of random background noise. Next the filtered signal is segmented into 256-sample frames with the overlap of 128 samples. Every short-time segment of the signal is windowed using Hanning function.

A perceptual cepstral analysis is used for feature extraction. 13th order Mel-scale cepstral features are extracted with 20 triangular filters used for processing of the magnitude spectrum covering entire frequency range of the signal (0 kHz–5.5 kHz). The first MFCC – average of the spectrum of the frame – is eliminated from the feature set.

The experimental investigation of the isolated word recognition system is tested by the use of 1.5 s duration recording, therefore, 128 feature vectors are extracted for each speech pattern.

The feature vectors are compared using DTW based algorithm. DTW algorithm is known for its capability to compare sequences (patterns) of different lengths and is widely used for isolated word recognition, signature recognition, speaker identification and string comparison. The idea of the DTW algorithm is to match two sequences and to evaluate their similarity in most coinciding points. The calculated similarity measure (called distance, dissimilarity or matching error) can be used for decision making about similarity of sequences. Having all distances between the unknown and reference patterns (stored in dictionary) enables to identify (recognize) the unknown pattern.

Every point of the grid corresponds to the pair of frames of comparable sequences and is tied with the inter-distance of the corresponding feature vectors. The goal of the DTW comparison is to find the path in the grid giving the minimal accumulated distance

$$D_{p,r} = \min \sum_{n=1}^{N} d\left(p_n, r_n\right), \tag{1}$$

where p_n is the vector set of the pattern features, r_n is the vector set of the reference pattern features and N is the number of references stored in the vocabulary.

Minimal distance represents the appropriate match of the sequences and distance calculation. The search of the minimal cost path (the selection of partial distances) is restricted by various conditions and constraints thus making comparison process consistent and valid in time.

Each word in the vocabulary of the recognition system has four related references, stored as a set of MFCC feature vectors. The recognition rate of the system is directly dependent to the stored reference set. In order to maximize the efficiency of the recognition system, the reference set should be close enough to the possible incoming word pattern and far enough from the incoming word patterns belonging to other words. Taking into account that the word pattern is highly dependent on the pronunciation and environment influence to the sound, the continuous renewal of the vocabulary should ensure the vocabulary references are up to date.

III. New Algorithm for Vocabulary Renewal

The vocabulary renewal algorithm uses two main rules to make the decision. The first rule measures the similarity of the newly recognized pattern to the same word references already stored in the memory. The second rule checks if addition of the new pattern will decrease the similarity between current word patterns and references, stored for the other words.

The novelty of this algorithm is in the original way of pattern comparison during the vocabulary renewal stage.

A. Reduction of the Dimensionality

The first and the second rules in the algorithm requires the comparison of references stored in the vocabulary (each reference should have the vector of distances to other references) and a newly recognized pattern (for each new pattern the distances to every reference should be estimated).

Taking into account, that the size of vocabulary is four times expanded because of four references stored for each word, the internal distance estimation will result in the 400×400 elements size matrix for the vocabulary of 100 different words. The arrival of the new pattern will require additional 400 comparisons to be performed in order to make a decision to update the vocabulary or not.

The previous implementation of the system [12] has DTW based matching module already implemented in FPGA thus it would be possible to use the same matching procedure for internal distance estimation. However the DTW procedure is highly computational intensive. To reduce the amount of computations an alternative way of pattern comparison is proposed in this paper.

The reduction of dimensionality for the MFCC vector sequence is performed by the use of self-organizing feature map (SOFM). The idea of such approach is to classify MFCC vectors into several classes in such manner, that the MFCC vectors estimated for the noisy silence at the beginning of the word and at the end of it would be classified into the same class.

The SOFM is selected because of the ability to converge into solution without the manual definition of the desired output. An unsupervised learning algorithm adapts the weights of the self-organizing feature map to maximally isolate the most different feature sets.

The maximum number of SOFM elements to be used for the Lithuanian isolated word vocabulary was experimentally found to be 6. The further increase of the number of map

elements will lead to classification of noisy silence windowed signal into more than one class.

B. Comparison of the Time-dependent MFCC Variations

The classified MFCC feature vectors give not only the grouping of the MFCC vectors but also provide the time-dependent distribution of the self-organizing feature classes among the time axis. The result of MFCC vector classification is shown in Fig. 1. The time axis in this figure is directly related to the MFCC feature vector number, estimated for 256 signal samples with short-time analysis window overlap by 128 samples.

The application of the classifier for different pronunciations of the same word has showed that the transition between neighbouring well expressed classes (when a sequence of successive MFCC vectors is assigned to the same class) might be different. As it is seen in Fig. 1, one or several intermediate classes during transition might be different but it is well seen that the time-depended structure of the MFCC classes is very similar.

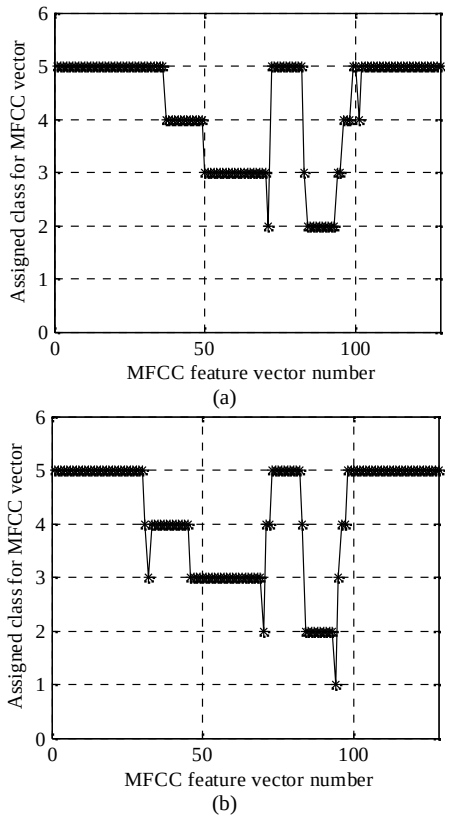

Fig. 1. Visualization of the MFCC feature vector classification result: (a) – first pronunciation; (b) – second pronunciation.

In order to estimate the similarity between two pronunciations of the same word the result of MFCC vector classification are treated as the sequence of symbols (a sequence of classes). Taking into account that some parts of the sequences, estimated for the same word are similar or identical we propose to apply the Needleman-Wunsch sequence matching algorithm [13] to estimate the similarity score of the two compared sequences. The highest received matching score indicates maximum similarity in the compared set of reference sequences.

The summary of the proposed algorithm is provided in Algorithm 1.

Algorithm 1. Vocabulary renewal for speech recognition system.

A. Reduction of the dimensionality for the vocabulary references:

1. Classification of reference MFCC feature vectors r_n using self-organizing feature map – estimation of the time-dependent MFCC feature variations vector $R_{\mathrm{SOM},n}$.

2. Application of Needleman-Wunsch sequence alignment algorithm to estimate the similarity between current reference sequence $R_{\mathrm{SOM},C}$ and every other reference sequence $R_{\mathrm{SOM},n}$ stored in the vocabulary.

B. Comparison of the newly recognized pattern to the references stored for the same word:

1. Classification of new pattern MFCC feature vectors p_n using self-organizing feature map – estimation of the time-dependent MFCC feature variations vector $P_{\mathrm{SOM},n}$.

2. Application of Needleman-Wunsch sequence alignment algorithm to estimate the similarity $d\left(P_{\mathrm{SOM},C}, R_{\mathrm{SOM},r}\right)$ between new pattern sequence $P_{\mathrm{SOM},C}$ and all four related reference sequences $R_{\mathrm{SOM},r}$ stored in the vocabulary and assigned to the recognized word.

IF $d\left(P_{\mathrm{SOM},C}, R_{\mathrm{SOM},r}\right) < \max\left\{d\left(R_{\mathrm{SOM},r*}, R_{\mathrm{SOM},r}\right)\right\}$, where $d\left(R_{\mathrm{SOM},r*}, R_{\mathrm{SOM},r}\right)$ is the similarity between same word references stored in the vocabulary.

THEN $P_{\mathrm{SOM},C}$ is selected as potential candidate to be selected for vocabulary update. Pair of sequences $R_{\mathrm{SOM},r*}$ and $R_{\mathrm{SOM},r}$ resulting $\max\left\{d\left(R_{\mathrm{SOM},r*}, R_{\mathrm{SOM},r}\right)\right\}$ is selected as potential candidates to be removed from the vocabulary.

ELSE wait for new pattern and return to part B.

ENDIF

C. Decision on the reference update

IF $P_{\mathrm{SOM},C}$ is not empty

IF $\max\left\{d\left(R_{\mathrm{SOM},r*}, R_{\mathrm{SOM},n}\right)\right\} < \max\left\{d\left(R_{\mathrm{SOM},r}, R_{\mathrm{SOM},n}\right)\right\}$

$R_{\mathrm{SOM},r}$ is selected as a potential candidate $R_{\mathrm{SOM},C}$ to be removed.

ELSE

$R_{\mathrm{SOM},r*}$ is selected as a potential candidate $R_{\mathrm{SOM},C}$ to be removed.

ENDIF

IF $\max\left\{d\left(P_{\mathrm{SOM},C}, R_{\mathrm{SOM},n}\right)\right\} < \max\left\{d\left(R_{\mathrm{SOM},C}, R_{\mathrm{SOM},n}\right)\right\}$

$P_{\mathrm{SOM},C}$ is selected as a new vocabulary reference instead $R_{\mathrm{SOM},C}$.

ELSE wait for new pattern and return to part B.

ENDIF

ELSE wait for new pattern and return to part B.

ENDIF

IV. RESULTS OF THE EXPERIMENTAL INVESTIGATION

The experimental investigation was performed on the Lithuanian speech isolated word recordings. 100 different words were used for the vocabulary and four different pronunciations for each word were stored as the references.

The performance of proposed vocabulary renewal solution was measured by the analysis of four closest references to the newly arrived pattern, estimated using DTW based matching algorithm. If the recognition of the new pattern (new pronunciation of the word with references stored in the vocabulary) is possible with every reference assigned to the same word – four closest references in the vocabulary are assigned to the same word, then the set of references could be treated as optimal.

In order to affirm that the proposed vocabulary renewal algorithm is applicable the number of word recognitions should not decrease with all four same word references standing as the four best solutions accordingly to DTW. Also it is needed to affirm, that the update of the vocabulary does not reduce the recognition rate for the last several

examples.

TABLE I. NUMBER OF WRONG WORD REFERENCES IN A SET OF FOUR BEST DTW MATCHES.

Vocabulary Update Phase	Wrong Word References (from 1st to 4th)	Number of updated references (number of candidates)
Initial set of references	37 (2, 2, 6, 27)	–
	84 (3, 11, 24, 46)	–
Updated after 1st pronunciation	35 (2, 2, 5, 26)	18 (45)
	90 (3, 13, 25, 49)	26 (63)
Updated after 2nd pronunciation	36 (2, 1, 6, 27)	22 (58)
	90 (4, 12, 24, 50)	29 (62)
Updated after 3rd pronunciation	35 (2, 1, 7, 25)	70 (93)
	91 (3, 13, 24, 51)	31 (60)
Updated after 4th pronunciation	35 (2, 1, 7, 25)	59 (92)
	91 (3, 13, 24, 51)	31 (66)
Updated after 5th pronunciation	35 (2, 1, 7, 25)	30 (67)
	91 (3, 13, 24, 51)	17 (50)

The results of experimental investigation are given in Table I. Here the recognition results for two sets of pronunciations, provided by two different persons are compared. Each test set used for experimental investigation consists of 10 pronunciations of 100 different words. Four pronunciations are used as initial set of references in the vocabulary. The DTW comparison results, provided in the Table I, are received by matching the 10th pronunciation with the references in the vocabulary. The update of vocabulary references is performed according to Algorithm 1. For update, the rest of the pronunciations (from 5th to 9th) were used. The first number in the second column (see Table I) shows the total number of wrongly assigned references. The first numbers in the brackets shows the wrongly matched references (from 100 possible), from those which gave the minimal difference according to DTW. Following numbers shows, how many wrong matches would be if we use second minimal difference, third, etc.

Looking at the third column in Table I it is seen, that different pronunciation provides different number of updated references and actually is more dependent on the differences in pronunciation than on the number of previously made updates.

Additional experimental investigation was performed in order to check if the continuous update of the vocabulary references does not decrease the recognition rate of the last five pronunciations. The results has shown, that the updated vocabulary still remains capable to recognize the isolated words with higher rate, comparing to the situation before reference update. The four phases of vocabulary update still gave the 100 % recognition (0, 4, 6, 21), whereas initially, the isolated word recognition system were not able to correctly recognize 3 patterns (3, 13, 24, 51). The updated vocabulary also showed better performance for the 6th pronunciation (0, 0, 6, 20), 7th (0, 5, 11, 18) and 8th (1, 2, 8, 18). The 9th pronunciations were not tested because the last update of the vocabulary was made accordingly to 9th set of patterns.

V. CONCLUSIONS

The application of the self-organizing feature map to the sequence of the MFCC feature vectors, used for isolated word speech recognition system gives ability to receive a time-dependent sequence of the MFCC feature sets classified into one from several classes. The alignment of received sequences of classes could be performed by the use of Needleman-Wunsch sequence matching algorithm giving a unique maximal matching.

The application of continuous vocabulary update does not degrade the recognition rate of the unknown pattern, received from new pronunciation, but highly improves the recognition of the last used patterns. It shows that the proposed algorithm is able to adapt to the previous examples without need to increase the number of references stored.

REFERENCES

[1] G. Tamulevicius, V. Arminas, E. Ivanovas, D. Navakauskas, "Hardware accelerated FPGA implementation of Lithuanian isolated word recognition system", *Elektronika ir Elektrotechnika*, vol. 99, no. 3, pp. 57–62, 2010.

[2] B. King, P. Smaragdis, G. Mysore, "Noise-robust dynamic time warping using PLCA features", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012, pp. 1973–1976. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2012.6288293

[3] S. Xihao, Y. Miyanaga, "Dynamic time warping for speech recognition with training part to reduce the computation", in *Int. Symposium on Signals, Circuits and Systems (ISSCS 2013)*, 2013, pp. 1–4. [Online]. Available: http://dx.doi.org/10.1109/ISSCS.2013.6651195

[4] J. Peltola, J. Plomp, T. Seppanen, "A dictionary-adaptive speech driven user interface for a distributed multimedia platform", in *Proc. 25th EUROMICRO*, 1999, pp. 326–332.

[5] J. Chien, C. Chueh, "Dirichlet class language models for speech recognition", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 482–495, 2011. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2050717

[6] H. Li, T. Zhang, R. Qiu, L. Ma, "Grammar-based semi-supervised incremental learning in automatic speech recognition and labelling", *Energy Procedia*, vol. 17, no. 1, pp. 1843–1849, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.egypro.2012.02.321

[7] D. Biswajit, M. Sandipan, M. Pabitra, B. Anupam, "Aging speech recognition with speaker adaptation techniques: Study on medium vocabulary continuous Bengali speech", *Pattern Recognition Letters*, vol. 34, no. 3, pp. 335–343, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2012.10.029

[8] L. Henderson, A. Weighall, G. Gaskell, "Learning new vocabulary during childhood: Effects of semantic training on lexical consolidation and integration", *Journal of Experimental Child Psychology*, vol. 116, no. 3, pp. 572–592, 2013. http://dx.doi.org/10.1016/j.jecp.2013.07.004

[9] T. Sledevic, A. Serackis, G. Tamulevicius, D. Navakauskas, "Evaluation of features extraction algorithms for a real-time isolated word recognition system", *Int. Journal of Electrical, Electronic Science and Engineering*, vol. 84, pp. 290–294, 2013.

[10] V. Tyagi, C. Wellekens, "On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 529–532.

[11] J. Manikandan, B. Venkataramani, "Design of a real time automatic speech recognition system using modified one against all SVM classifier", *Microprocessors and Microsystems*, vol. 35, no. 6, pp. 568–578, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.micpro.2011.06.002

[12] T. Sledevic, D. Navakauskas, "FPGA-based fast Lithuanian isolated word recognition system", in *IEEE EUROCON*, 2013, pp. 1630–1636.

[13] S. B. Needleman, Ch. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970. [Online]. Available: http://dx.doi.org/10.1016/0022-2836(70)90057-4