THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 16, Number 1/2015, pp. 103-109

VOCABULARY DISTANCE MATRIX ANALYSIS-BASED REFERENCE TEMPLATE UPDATE TECHNIQUE

Gintautas TAMULEVICIUS[1], Arturas SERACKIS[2], Tomyslav SLEDEVIC[2], Dalius NAVAKAUSKAS[2]

[1] Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania
[2] Department of Electronic Systems, Vilnius Gediminas Technical University, Vilnius, Lithuania
Corresponding author: Gintautas TAMULEVICIUS, E-mail: gintautas.tamulevicius@mii.vu.lt

Intra-reference and inter-reference distances can be used for evaluation and comparison of reference templates in template based speech recognition. Hence decision on quality of reference template can be done and the reference set can be updated if necessary. In this paper a new reference set update technique based on analysis of vocabulary distance matrix is proposed. Reference templates with small intra-reference and inter-reference distance values are substituted with candidate patterns giving bigger distance values. The substitution can be performed in supervised or unsupervised manner. The technique was tested for Dynamic time warping based isolated word recognizer. The reference substitution using proposed technique increased the intra-reference and inter-reference distances by 6.8 % and 13.6 % respectively.

Key words: Human computer interaction, Automatic speech recognition, Pattern matching.

1. INTRODUCTION

Changing acoustical conditions inevitably deteriorate speech recognition rate. Variable speaking style and rate, individual speaker parameters, background noise level, differences in input channels are the main acoustical factors influencing and hence defining the actual accuracy of the speech recognizer [1, 2].

The mainstream solution to this problem is the adaptation of speech recognizer. During the adaptation process the reference models (or templates) of the recognizer are modified (adapted or updated) in accordance with the data obtained under new and thus unknown acoustical circumstances. Let's name this data – adaptation data. Usually the amount of required data is much smaller in comparison with the amount of utilized training data. This is the main advantage of adaptation over the additional training (or even retraining) of the recognizer.

The main goal of reference modification is the adaptation to a new speaker. This allows reducing recognition errors in speaker-independent systems. Adaptation to noisy environment and to present speaker in order to improve recognition rate are explored, too.

In this paper we present a new technique for reference template update using speech recognizer's vocabulary based distance matrix. The proposed update of templates can be performed in supervised or unsupervised manner.

The paper is organized as follows. In Section II currently employed reference template adaptation and update techniques are reviewed. In Section III new reference template update technique together with two possible to use criteria are proposed. In Section IV experimental results on speech recognizers' reference template update are summarized. General conclusions are given in Section V.

2. TEMPLATE ADAPTATION TECHNIQUES USED FOR SPEECH RECOGNITION

The update of reference models can be performed in supervised or unsupervised manner [3, 4]. In supervised case some confirmation on correct or false recognition is needed. Considering the recognition

result the appropriate reference model is updated positively or negatively (making more "distant" from input speech pattern) thus making the model more robust to the speaker or noise. Nowadays most used adaptation techniques are supervised. Unsupervised adaptation should use some automated decision making scheme that allows to reject improper speech patterns. This way of processing is prone to erroneous adaptation and should be used carefully.

The reference models can be updated in batch or on-line mode [5]. Batch-mode adaptation is performed when all proper candidate utterances are collected during some time period and are used in predefined moment. On-line adaptation can be performed straight after appropriate and confirmed candidate pattern is obtained.

The adaptation technique for the speech recognizer is determined by used speech recognition approach. Two main approaches are used for different recognition tasks.

For large vocabulary speech recognition Hidden Markov model approach (HMM) is used [6, 7]. It is a statistically-based approach using statistical modelling of speech patterns (words, phonemes, etc.). Therefore, all HMM models are adapted (updated) statistically: Maximum Likelihood Linear Regression [8] and Maximum a Posteriori [9] approaches are widely used for this purpose. More recent techniques include fuzzy logic [10], linear interpolation of models [11], decision trees [12]. Comprehensive list of statistical adaptation techniques can be found in [13].

Small vocabulary speech recognition is implemented using pattern comparison techniques. The main idea of pattern comparison based speech recognition is the comparison of unknown speech pattern with reference templates. The most similar template is chosen as the equivalent to unknown speech pattern. The Dynamic Time Warping algorithm [14] is used for pattern comparison mostly [15, 16].

Commonly, because of the template is stored for every speech pattern the adaptation of such recognizer is performed by replacing the original template with new one. In [3] the averaging procedure (with DTW alignment) is applied for the correct reference and input patterns. Every time the averaging is performed using weighting scheme so that the modified template is the average of the initial template and all input patterns used for adaptation. The classification error minimizing discriminative training algorithm is proposed for candidate patterns selection in [17]. In [18] the newly adapted reference template is assessed using a distance between the adapted reference and the current reference template. Besides, the distance between adapted reference and other reference templates is considered additionally.

In summary, the main criterion of selecting templates to be substituted (or updated in some manner) is the distance (dissimilarity degree) between candidate pattern and the current reference template.

3. NEW REFERENCE TEMPLATE UPDATE TECHNIQUE

The pattern comparison and decision making in speech recognition is based on objective distance that gives a numerical value of similarity. The distance is calculated for every reference pattern thus obtaining a sequence of distance values. The decision on match is carried out using the minimum criterion and some threshold value (in order to reject possibly false recognition).

The set of reference templates for the recognition system can be characterized using entire vocabulary distance matrix. The distance matrix can be obtained by pairwise comparison of reference templates. The considered structure of similarity matrix is given in Figure 1.

Each matrix element represents the distance between pair of reference templates. Let's name the distance between different reference templates as inter-reference distance. The term "intra-reference distance" will be used to denote the distance between different examples of the same reference template.

Suppose K examples of N reference templates are collected. Thus the matrix element $d_{21,11}$ shows the distance between the first example of the second reference pattern and the first example of the first reference pattern. The diagonal of the matrix corresponds to distances between the same examples of reference templates hence it will be filled with zero values.

		r_1			...	r_N		
		r_{11}	...	r_{1K}		r_{N1}	...	r_{NK}
r_1	r_{11}	$d_{11,11}$		$d_{11,1K}$		$d_{11,N1}$		$d_{11,NK}$
	⋮		...		...		...	
	r_{1K}	$d_{1K,11}$		$d_{1K,1K}$		$d_{1K,N1}$		$d_{1K,NK}$
⋮			...		...		...	
r_N	r_{N1}	$d_{N1,11}$		$d_{N1,1K}$		$d_{N1,N1}$		$d_{N1,NK}$
	⋮		...		...		...	
	r_{NK}	$d_{NK,11}$		$d_{NK,1K}$		$d_{NK,N1}$		$d_{NK,NK}$

Fig. 1 The possible structure of vocabulary distance matrix

The distance matrix can be symmetric (sometimes referred as Euclidean distance matrix) or asymmetric. The asymmetric distance matrix usually is attained in the case of asymmetric pattern comparison. For example Dynamic Time Warping algorithm with asymmetric local constraints (ensuring temporal consistency of comparison process) gives an asymmetric distance matrix [2]. In this case the whole matrix should be taken into account.

The main requirements for reference templates are generality and separability. Generality means ability to cover various speakers (in speaker-independent recognition case) or varying pronunciation (in speaker-dependent case). In terms of distance matrix higher generality degree means higher values for intra-reference distances $d_{i,i}$, for $i = 1, \ldots, N$.

Separability is determined by the difference of distinct reference templates. Higher difference means higher discriminative power of references thus lower recognition error rate. The separability is reflected by inter-reference distances $d_{i,j}$, when $i, j = 1, \ldots, N$ and $i \neq j$. Thus higher inter-reference distance value means higher discriminative power of the reference template. When the minimal inter-reference distance is smaller than minimal intra-reference distance the recognition error is obtained according to minimum criterion based decision scheme.

Accordingly, following requirements for reference template can be formulated:

- Bigger minimal intra-distance;
- Bigger minimal inter-distance;
- Minimal inter-distance bigger than minimal intra-distance.

According to outlined requirements reference templates can be compared and decision on their quality can be done. In general, the reference with bigger inter- and intra-distances is much more desirable than the reference with smaller ones.

By applying above mentioned requirements to input patterns the reference update criterion can be implemented. The input pattern (confirmed as recognized correctly) which overtakes the current reference template agreeably to these requirements should be considered as reference with higher degree of generality and separability. Substitution of the current reference template into new-found pattern should ensure more reliable and robust comparison and thus recognition process.

Let us compose expressions for reference template substitution. Suppose that candidate pattern P which was correctly matched with reference template r_{pk} is available. The template r_{pk} has to be substituted with pattern P only if both following conditions are satisfied:

- The pattern gives more generality than r_{pk}

$$\min_k d_{P,\,pk} > \min_{l,\,l \neq k} d_{pl,\,pk}, \quad k, l = 1, \ldots, K. \tag{1}$$

- The pattern gives more separability than r_{pk}

$$\min_{i,\, i\neq p} d_{P,\, ik} > \min_{i,\, i\neq p,\, l\neq k} d_{pk,\, il}, \quad i=1,\ldots,N, \quad k,l=1,\ldots,K; \tag{2}$$

$$\min_{i,\, i\neq p} d_{P,\, ik} > \min_{k} d_{P,\, pk}, \quad i=1,\ldots,N, \quad k=1,\ldots,K. \tag{3}$$

Reference pattern r_{pk} which meets conditions defined in (1)–(3) should be substituted with the pattern *P* if

$$\arg\min_{k} d_{P,\, pk} = \arg\min_{i,\, i\neq p} d_{P,\, ik}, \quad i=1,\ldots,N, \quad k=1,\ldots,K. \tag{4}$$

The reference substitution algorithm is depicted in Figure 2.

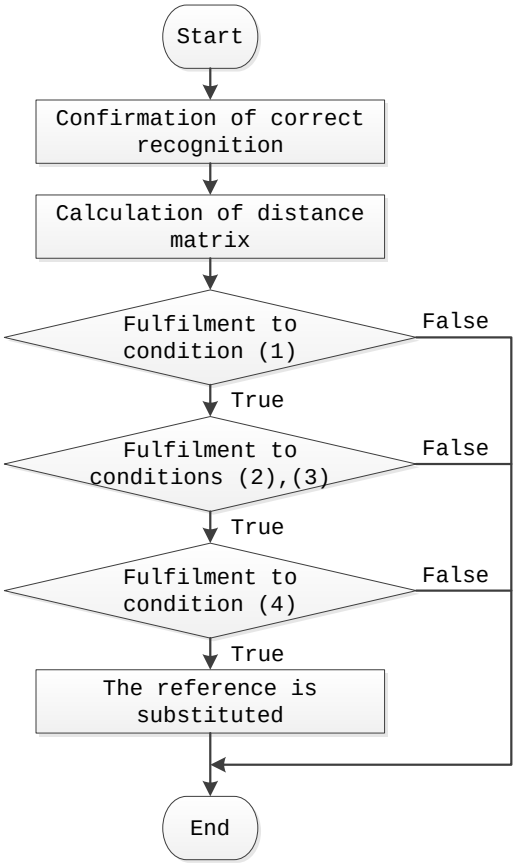

Fig. 2 Reference substitution algorithm (conditions in algorithms correspond to given above mathematical expressions).

The reference substitution is performed removing the obsolete reference pattern and storing the new one. In terms of distance matrix it would require recalculation of distance matrix particular row and column.

In case of symmetrical distance matrix data rows or columns can be taken for analysis. In case of asymmetric matrix the decision according to recognizer configuration should be made. If recognizer operates distances allocated in matrix rows they should be taken for analysis. And according to data row analysis results, the column data should be recalculated, too.

The reference substitution can be performed in unsupervised manner. In this case some threshold value for distance $d_{P,pk}$ is needed in order to make automated decision about correctness of the recognition result.

After the reference substitution, the new version of distance matrix is obtained. In order to reveal the effect of substitution, some objective criterion for comparison of different versions of distance matrix needs to be developed.

The initial distance matrix D and the updated matrix version D^+ can be compared using various methods, e. g., Mantel's test, cross-correlation based measures. In the following matrix norm-like static and dynamic comparison criterions are proposed.

The minimal distance (MD) criterion can be expressed by:

$$C_{\mathrm{MD}} = \min_{i,\, j,\, n,\, m} d_{in,\, jm}. \tag{5}$$

The MD criterion can be calculated for intra-reference ($i = j$) and for inter-reference ($i \neq j$) cases. Small minimal intra-reference and big minimal inter-reference distance values will show high degree of separability of reference templates.

The averaged change of C_{MD} values calculated from initial D and updated D^+ distance matrices will show *the average change of minimal distance* (MDC) values and will characterize the reference substitution process. Thus the change of minimal value will be

$$C_{\mathrm{MDC}} = \frac{1}{KN}\left(\min_{i,\, j} d_{i,\, j} - \min_{i,\, j} d^{+}_{i,\, j}\right). \tag{6}$$

Again, the average change of minimal distance criterion can be used for estimation of intra-reference ($i = j$) and inter-reference ($i \neq j$) distances. Small intra-reference C_{MDC} value and big inter-reference C_{MDC} value will point to successful reference substitution process.

4. EXPERIMENTAL RESULTS

The proposed reference substitution technique was experimentally tested in speaker-dependent isolated word recognition task. The DTW based isolated word recognizer was used for this purpose [19]. The 12th order Perceptual Linear Predictive coding analysis was selected for acoustical analysis as one of the mostly used in modern speech recognition tasks. The speech recognizer was implemented and the whole experiment was executed in MATLAB environment.

Ten Lithuanian words (meaning the numbers from 0 to 9) were selected for the vocabulary of the recognizer. Records of two speakers (male and female) were used. Every word was pronounced 9 times by each speaker thus obtaining the set of 180 utterances (2 speakers × 9 pronunciations × 10 words). First two pronunciations of words were assigned for initial training of the recognizer. Next 6 pronunciations were used for reference set update, i. e., were used in reference substitution process. The reference set update process was performed in supervised manner. The last set (denoted as a test set) of utterances was used for evaluation of recognition rate before and after the update of reference set.

Firstly the minimal intra-reference and inter-reference distances C_{MD} were analysed before and after the reference set update. The results are given in Table 1.

Table 1 The change of intra-reference and inter-reference distances

	Before update		**After update**		**Change, %**	
Speaker	Minimal intra-refernce distance	Minimal inter-refernce distance	Minimal intra-refernce distance	Minimal inter-refernce distance	Minimal intra-refernce distance	Minimal inter-refernce distance
Female	0.088	0.152	0.088	0.163	0.0	7.2
Male	0.064	0.129	0.732	0.155	13.6	19.9

From the table data becomes evident that the reference set update gave average 6.8 % increase of intra-reference and average 13.6 % increase of inter-reference update. In case of female record set 3 references

were substituted, in case of male record set 7 references were substituted. This explains the larger distance changes for male utterances.

Further experiments with larger initial reference sets (with more templates per word) has shown that update of larger reference sets is less probable and the change of minimal distances decreases to zero. Thus the proposed reference substitution technique should be applied for smaller reference sets in order to enhance their generality and separability properties.

The results of MDC criterion calculation for inter-reference distance are given in Fig. 3.

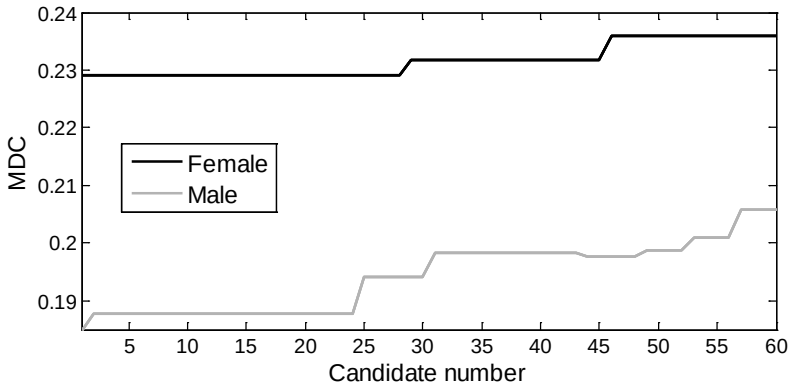

Fig. 3 MDC criterion dependence on candidate number

As it is seen the update for male speaker reference set was more frequent and the dynamics of minimal inter-reference distance was more intensive. The average change of MDC criterion was around 11 % for male speaker case and 3 % for female speaker case. Analysis of recorded data has shown that pronunciation and acoustical quality of female speaker records was higher. Thus update level and dynamics of distance matrix values is conditioned by initial reference set. The smaller intra-reference and inter-reference distances are in distance matrix the more intensive reference set update process could be expected.

In case of male speaker negative change of MDC value (with 45th candidate utterance) was attained however the overall change still remained positive.

The obtained speech recognition rate (determined using test set) was 100 % for all versions of reference sets both for male and female record sets. In case of lower recognition rates for initial reference sets some increase of recognition after the reference sets were updated could be expected.

5. CONCLUSIONS

The paper presents a new reference set update technique for isolated word recognition system. The initial references are substituted if their intra-reference and inter-reference distances are smaller than the distances provided by candidate utterance. The following conclusions can be stated:

- The reference template can be substituted considering its intra-reference and inter-reference distances. The substitution of references using this technique increased the average inter-reference distance by 13.6 %.
- The proposed technique is more applicable for small reference sets with a few reference templates per word. The update of larger reference sets is less probable.
- The final results of reference set update are speaker-dependent and the overall improvement of set depends on initial set.

ACKNOWLEDGMENTS

This research was funded by a grant (No. MIP-092/2012) from the Research Council of Lithuania.

REFERENCES

1. T. VIRTANEN, R. SINGH, B. RAJ, *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons, 2013.
2. T. SLEDEVIC, G. TAMULEVICIUS, D. NAVAKAUSKAS, *Upgrading FPGA implementation of isolated word recognition system for a real-time operation*, Elektronika ir Elektrotechnika, **19**, pp. 123–128, 2013.
3. F. R. MCINNES, M. A. JACK, J. LAVER, *Template adaptation in an isolated word-recognition system*, IEE Proc. Com., Speech and Vision, **136**, pp. 119–126, 1989.
4. K. SHINODA, *Speaker adaptation techniques for speech recognition using probabilistic models*, Electronics and Communications in Japan (Part III: Fundamental Electronic Science), **88**, pp. 25–42, 2005.
5. G. ZAVALIAGKOS, *Batch, incremental and instantaneous adaption techniques for speech recognition*, in Proc. of Int. Conf. Acoustics, Speech, and Signal Processing ICASSP-95, **1**, pp. 676–679, 1995.
6. L. RABINER, B.-H. JUANG, *Fundamentals of speech recognition*. New Jersey: Prentice-Hall, 1993.
7. M. A. YUSNITA, M. P. PAULRAJ, S. YAACOB, S. A. BAKAR, A. SAIDATUL, A. N. ABDULLAH, *Phoneme-based or isolated-word modeling speech recognition system? An overview*, in IEEE 7th Int. Colloquium on Signal Processing and its Applications (CSPA), pp. 304–309, 2011.
8. C. J. LEGGETTER, P. C. WOODLAND, *Maximum likelihood linear regression for speaker adaptation of continuous density Markov models*, Computer Speech & Language, **9**, pp. 171–185, 1995.
9. C.-H. LEE, C.-H. LIN, B.-H. JUANG, *A study on speaker adaptation of continuous density HMM parameters*, in Proc. of Int. Conf. Acoustics, Speech, and Signal Processing ICASSP-90, **1**, pp. 145–148, 1990.
10. I.-J. DING, *Fuzzy logic based control for VFS speaker adaptation*, Int. J. of Fuzzy Systems, **10**, pp. 292–297, 2008.
11. W.-X. TENG, G. GRAVIER, F. BIMBOT, F. SOUFFLET, *Rapid speaker adaptation by reference model interpolation*, in INTERSPEECH 2007, pp. 258–261, 2007.
12. T. SCHULTZ, A. WAIBEL, *Polyphone decision tree specialization for language adaption*, in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing ICASSP '00, **3**, pp. 1707–1710, 2000.
13. J. R. BELLEGARDA, *Statistical language model adaptation: review and perspectives*, Speech Communication, **42**, pp. 93–108, 2007.
14. H. SAKOE, S. CHIBA, *Dynamic programming algorithm optimization for spoken word recognition*, in IEEE Trans. on Acoustics, Speech and Signal Processing, **26**, pp. 43–49, 1978.
15. J. M. VALENCIA–RAMIREZ, A. CAMARENA–IBARROLA, *On aligning techniques, feature extraction and distance measures for isolated word recognition*, in Proc. IEEE Int. Conf. Power, Electronics and Computing, pp. 1–6, 2013.
16. S. C. SAJJAN, C. VIJAYA, *Comparison of DTW and HMM for isolated word recognition*, in Int. Conf. Pattern Recognition, Informatics and Medical Engineering (PRIME), pp. 466–470, 2012.
17. Y. LIU, Y.-C. LEE, H.-H. CHEN, G.-ZH. SUN, *Adaptive template method for speech recognition*, in Proc. IEEE-SP Workshop Neural Networks for Signal Processing, pp. 103–110, 1992.
18. S. DOBLER, R. SCHLEIFER, A. KIESSLING, R. BRUCKNER, *Reference templates adaptation for speech recognition*, European patent 1 205 906 A1, November 7, 2000.
19. A. SERACKIS, T. SLEDEVIC, G. TAMULEVICIUS, D. NAVAKAUSKAS, *Word recognition acceleration by double random seed matching in perceptual cepstrum error space*, in Proc. European Modelling Symposium, pp. 274–279, 2013.

Received July, 2014