A COMPARISONAL STUDY OF THE MULTI-LAYER KOHONEN SELF-ORGANIZING FEATURE MAPS FOR SPOKEN LANGUAGE IDENTIFICATION

[1,2]*Liang Wang* [1,2]*Eliathamby Ambikairajah* [2,1]*Eric H.C. Choi*

[1]School of Electrical Engineering and Telecommunications
The University of New South Wales
Sydney, NSW 2052, Australia
[2]ATP Research Laboratory
National ICT Australia
Sydney, NSW 1435, Australia

ABSTRACT

Our previous research indicates that the multi-layer Kohonen self-organizing feature map (MLKSFM) gives a promising performance for spoken language identification (LID). In this paper, we enhance this approach in two distinct ways. Firstly, by considering the phase information, we propose a new type of feature vector which combines the modified group delay function (MODGDF) and the traditional MFCC. Secondly, we propose a hierarchical structure of the MLKSFM, in which the pre-classification is performed at the lower level MLKSFM and the final language identification is performed at the top level MLKSFM. For the OGI-TS speech corpus, the best LID rate is achieved at 87.3% for the 45-sec test speech utterances by using the hierarchical MLKSFM with 4 classes pre-classified at the lower level MLKSFM. For the 10-sec test speech utterances, the best LID rated is achieved at 60.0% by using the non-hierarchical MLKSFM LID system.

Index Terms— Language identification, hierarchical multi-layer Kohonen self-organizing feature map, modified group delay function

1. INTRODUCTION

The goal of automatic language identification (LID) is to identify the language spoken in a particular utterance. Over the past decades, many approaches have been proposed to deal with the LID task [1][2][3]. The most well known methods for the LID task include the PPRLM (Parallel Phone Recognition and Language Modeling) based [1][2] and the GMM-UBM (Gaussian Mixture Model and Universal Background Model) based [3] LID systems.

For a system to yield high performance in language identification, two important properties must be comprised: the ability to extract sufficient information for the speech signal, and the ability to realize complex decision regions in the feature vector space [1][3].

The Kohonen self-organizing feature map [4][5][6] is a topology-preserving map from a high-dimensional input descriptor space to a lower dimensional grid or plane. Previous research indicates that the KSFM based systems require significantly less training time compared with other neural network systems [4][5]. Our recent research [7] has shown that by using segment-based input feature vectors, KSFM based language identification system is capable of achieving a similar identification rate compared with the phone-based language identification systems, but requires less training time and no phone labeling of training data.

Current state of the art language identification systems use speech features derived from the Fourier transform of spectral magnitude, like Mel-frequency cepstral coefficients (MFCC) and their derivatives. Thus the entire information in the speech signal may not be captured since the phase spectrum is ignored. In this paper, the LID rate of using the phase information on the variations of Kohonen self-organizing feature map (KSFM) based LID system is also examined. As it is suggested in [8], the modified group delay features (MODGDF) are used.

Our previous study [9] also indicated that, when we combined the tonal and non-tonal language pre-classification with the traditional phone based language identification system (PPRLM), the language identification rate was increased significantly, and also the computation time was largely reduced. In our PPRLM with tonal and non-tonal language pre-classification system, all languages were firstly pre-classified as tonal or non-tonal language. The final identification was then performed by using different PPRLM systems for tonal and non-tonal language separately.

Therefore in this paper we propose a novel MLKSFM language identification system by using a hierarchical structure with pre-classification. All languages are firstly pre-classified into several classes by the lower level of the MLKSFM based on unsupervised self-learning, and then the

final identification is performed by the top level of the MLKSFM.

2. THE MODIFIED GROUP DELAY FEATURE

Given a discrete-time signal **x[n]**, let **X(ω)** be its Fourier transform, then we have

$$X(\omega) = |X(\omega)| e^{j\theta(\omega)} \tag{1}$$

$$\log X(\omega) = \log(|X(\omega)|) + j\theta(\omega) \tag{2}$$

$$\tau(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{3}$$

where **θ(ω)** is the unwrapped phase function and **τ(ω)** is the group delay function. The group delay function can also be computed from the speech signal as in [8] using

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{4}$$

where the subscripts **R** and **I** denote the real and imaginary parts of the Fourier transform. **X(ω)** and **Y(ω)** are the Fourier transforms of **x[n]** and **n·x[n]** respectively. Since the spiky nature of the group delay spectrum, a 6-order median filter **H(ω)** is applied to the denominator term **|X(ω)|**. Thus the modified group delay feature (MODGDF) is defined as

$$\tilde{\tau}(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2 \, H(\omega)} \tag{5}$$

To convert the MODGDF to some meaningful forms and also be able to be concatenated with other cepstral features such as MFCC, the MODGDF is converted to cepstral using the discrete cosine transform (DCT). Delta and acceleration parameters for the MODGDF can also be defined in a manner similar to that of the delta and acceleration parameters of MFCC.

3. MULTI-LAYER KOHONEN SELF-ORGANIZING FEATURE MAP FOR LANGUAGE IDENTIFICATION

In this section, the single-layer KSFM and the multi-layer KSFM LID systems will be briefly introduced. A better explanation of these two LID systems can be found in [7]. The novel hierarchical structure of MLKSFM LID system is then described in Section 3.3.

3.1 Single-layer KSFM for Language Identification

A single-layer KSFM with a hexagonal lattice for the language identification task is shown in Fig 1.

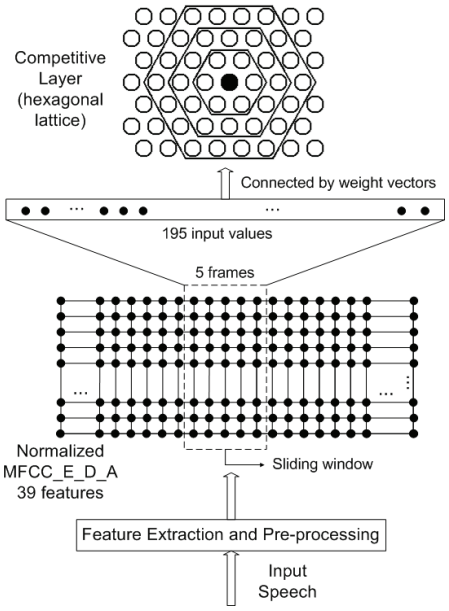

Fig. 1. The network structure of segment-based KSFM for language identification.

The basic feature vector contains 12 Mel-frequency cepstral coefficients (MFCC) and the log-energy. Additionally, the delta and acceleration coefficients are appended which results in a 39-dimension feature vector. Then the histogram equalization (HEQ) [10] is applied to all the feature vectors. In order to keep temporal information in the speech signal, we define a 5-frame window which moves over the whole speech utterance. By shifting one frame between windows, each window position yields a training vector of 195 dimensions (5 frames * 39 MFCC). Each of the 195-dimension segment-based training feature vectors is labeled for indicating its language identity.

For training the competitive neural layer, we generate a single training file containing the entire segment-based, normalized MFCC features from all the training speech utterances. During training, for each 195-diminsion input feature vector, a best-matched neural unit in the competitive layer is firstly selected. If the best match for the single input feature vector **x** is found at neuron **C**, then we have

$$\| x - \underline{w}_C \| = \min_j \| x - \underline{w}_j \| \tag{6}$$

where $\underline{\mathbf{w}}_\mathbf{j} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{195})$ is the weight vector (indexed by **j**) for each unit in the competitive layer, and || . || indicates the Euclidean norm.

If **n** is used to denote a discrete time index, then the weight vector is updated according to

$$\underline{w}_j(n+1) = \begin{cases} \underline{w}_j(n) + a(n)(x(n) - \underline{w}_j(n)), for\ j \in N_C \\ \underline{w}_j(n), \qquad\qquad otherwise \end{cases} \quad (7)$$

where **a(n)** is a positive constant that decays with time, and $\mathbf{N_C}$ defines a topological neighborhood around the best matched neuron unit **C**, which also decays with time.

Ideally on the completion of learning, each neural unit in the competitive layer should be made sensitive to only a certain language. In this case, the weight vector $\underline{\mathbf{w}}_\mathbf{j}$ for each unit is the representative vector of a certain language and the unit is given a label as the correspondent language.

During evaluation, for each unknown testing utterance **X**, the set of segmented, normalized feature vectors are first calculated as $\mathbf{X = \{ x_1, x_2, ..., x_t, ..., x_T \}}$. For each feature vector $\mathbf{x_t}$ the best match is found from each of the neural units in the competitive layer with a corresponding weight vector $\underline{\mathbf{w}}_\mathbf{j}$. The label **i** in $\underline{\mathbf{w}}_\mathbf{j}$ is added to the corresponding feature vector $\mathbf{x_t}$, where $\mathbf{i \in \Lambda, \Lambda = \{ 1, 2, \ldots, M \}}$ and **M** is the number of target languages. The final identification is performed by using a voting function:

$$\Phi(X) = \arg\max N(i \mid X), \quad \mathbf{i \in \Lambda} \quad (8)$$

where

$$N(i \mid X) = \sum_{t=1}^{T} \tau(x_t \in i) \quad (9)$$

is the number of votes for each language **i, i** $\in \mathbf{\Lambda}$, and

$$\tau(x_t \in i) = \begin{cases} 1, \text{if the unit for language } i \text{ is the best matched} \\ 0, \text{otherwise} \end{cases} \quad (10)$$

3.2 Multi-layer KSFM

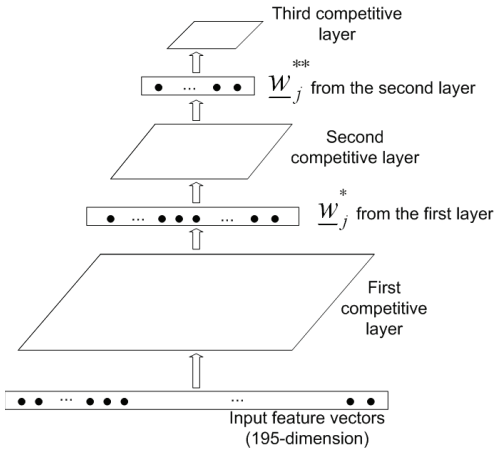

Fig. 2. The structure of multi-layer KSFM

Our previous research [7] indicates that for the language identification task, where a frame-based feature vector with very high dimensionality is used, the single-layer KSFM is not able to perform well. In such a situation the multi-layer KSFM (Fig. 2) can be of use.

The multi-layer KSFM (MLKSFM) used for language identification task is organized as a pyramidal structure consisting of multiple layers of single-layer KSFMs [7][11][12]. The number of neurons in each competitive layer decreases at each successive level. The input data arrives at the first layer and information is fed forward to higher layers. The weight vectors in each layer are converted into the input for the next layer. Thus in the higher layer of MLKSFM, each weight vector represents a higher level of abstraction of the input data.

During the training session, each labeled training vector activates one neural unit in the first competitive layer. The corresponding weight vector is then converted into the input training vector for the next layer by adding the same label in the training vector. Thus each training vector finally activates one neural unit in the top competitive layer.

The evaluation session of the MLKSFM is similar to the one used in the single-layer KSFM, and the final identification is performed by using the same voting function in equation (8).

3.3 Hierarchical Multi-layer KSFM

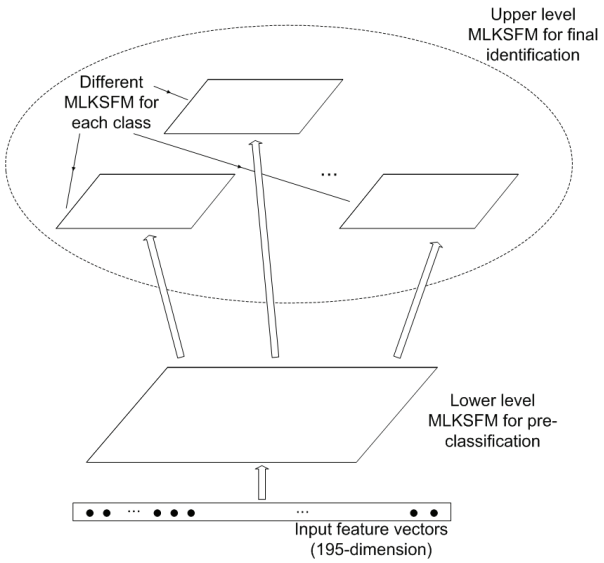

Fig. 3. The structure of hierarchical MLKSFM

With the multi-layer KSFM in hand, we extend our LID system to a hierarchical structure (Fig. 3). Firstly a lower level MLKSFM is used to perform a pre-classification. After the pre-classification all testing utterances will be classified into several classes and the classes are defined by checking the similarities of each activated neurons with the

corresponding languages on the top competitive layer from the lower level MLKSFM. For the pre-classification, the voting function is rewritten as

$$\Phi(X) = \arg\max N(i \mid X), \quad \mathbf{i} \in \underline{\mathbf{\Lambda}} \tag{11}$$

where $\underline{\mathbf{\Lambda}} = \{ \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \}$ and **N** is the number of language classes.

After the pre-classification, the final identification is performed by using different upper level MLKSFM in each pre-classified class for the unknown testing languages. The final identification is performed by using the same voting function in equation (8).

4. EXPERIMENT

We used the multi-layer KSFM system with the MFCC features only as a baseline system as it has better LID accuracy than the single-layer one. The OGI-TS speech corpus was used to perform the language identification task on the both the multi-layer KSFM and the hierarchical multi-layer KSFM systems. There are 11 languages in the corpus, namely English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Each speech utterance in the corpus was spoken by a unique speaker over a telephone channel. For testing the LID rate, the 45-sec "story-bt" utterances and the 10-sec "story-at" utterances were used. All testing utterances were unseen in training.

4.1. System Configuration

For the MLKSFM LID system, different types of topologies can be used in the competitive layer. For the local lattice structure, the hexagonal grid and the rectangular grid (Fig. 4) were used, while sheet and cylinder shapes (Fig. 5) were used to indicate the global map shape. Our previous research [7] indicated that the MLKSFM system with the sheet shaped map and the hexagonal lattice neighborhood relationship provided the best LID rate. Hence we only test our LID systems with this configuration in this experiment. For the first layer of the MLKSFM, the size was defined as 75*45. The sizes of the second and third layers were defined as 22*15 and 7*6 respectively.

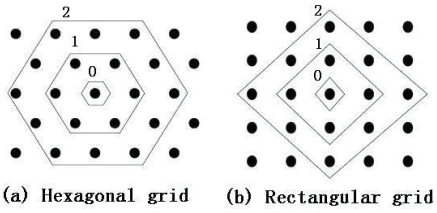

Fig. 4. Different lattice structures of the MLKSFM

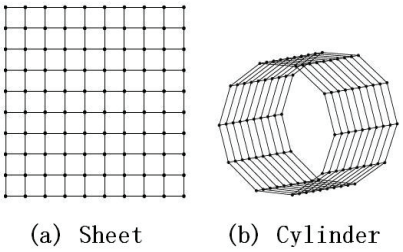

Fig. 5. Different map shapes of the MLKSFM

For the hierarchical MLKSFM system, the lower level pre-classification MLKSFM system and the upper level final identification MLKSFM system had the same structure and the same layer sizes as the MLKSFM LID system. Later in this paper we will explain how we define the classes in the lower level pre-classification MLKSFM system.

For both the MLKSFM and the hierarchical MLKSFM systems, the LID rates are also compared by using the MFCC only, MODGDF only and MFCC + MODGDF features.

For training the competitive layer, we used the sequential training algorithm instead of the batch training algorithm. The sequential training algorithm has a much lower memory requirement than batch training, at the cost of taking more time to compute.

4.2. Experimental Results

Table 1 shows the LID rates for the MLKSFM system by using different features. The LID rate is calculated as the number of correctly identified utterances out of all evaluation utterances. The best LID rate is obtained by using the MLKSFM system with the features derived by combining MFCC with MODGDF. The MODGDF alone gives the worst performance, but it still can be used for LID task, as the identification rates are 76.4% and 55.5% for the 45-sec and 10-sec testing utterances, respectively.

TABLE 1. COMPARISON OF LID RATES FOR MULTI-LAYER KSFM SYSTEM BY USING DIFFERENT FEATURES

Average length of test utterances	45-sec	10-sec
MLKSFM (MFCC)	78.2%	58.2%
MLKSFM (MODGDF)	76.4%	55.5%
MLKSFM (MFCC + MODGDF)	83.6%	60.0%

Fig. 6(a) plots the third competitive layer with the hexagonal grid and sheet shaped map in the MLKSFM after the training session. Each neural unit is activated by a particular language, and the corresponding label is added to that unit. It can be shown that most of the languages are able to form one or two clusters in the third competitive layer. More interestingly, some languages that activate adjacent clusters are those with the similar high-level language

features. For example, German and English are both stress-timed languages, and the neural units that are activated by these two languages are mostly located in the top left corner of the third competitive layer. Similar observations can also be found for two tonal languages—Mandarin and Vietnamese.

Fig. 6(b)-(d) show how we define the classes for the pre-classification by using the lower level MLKSFM in hierarchical MLKSFM. In this paper we only compare the results for the pre-classifying of 2 classes, 3 classes and 4 classes. In Fig. 6(b) we define 2 classes for the pre-classification (English, French, German, Korean, Spanish and Tamil are the first class while Hindi, Japanese, Farsi, Mandarin, and Vietnamese are the second class). In Fig. 6(c) we have 3 classes (English, German and Korean are the first class, French, Spanish and Tamil are the second class and Hindi, Japanese, Farsi, Mandarin, and Vietnamese are the third class). In Fig. 6(d) we have 4 classes defined by the lower level MLKSFM (English, German and Korean are the first class, French, Spanish and Tamil are the second class, Japanese and Farsi are the third class, and Hindi, Mandarin, and Vietnamese are the fourth class).

Table 2 compares the LID rate of the MLKSFM and the hierarchical MLKSFM by using different speech features. In this experiment we also perform the language identification by using separate MLKSFMs for different languages (thus we have 11 MLKSFMs for the 11 language classes in pre-classification). For the 10-sec speech utterances, the best LID rate is obtained by using the MLKSFM. While for 45-sec speech utterances, the best LID rate is obtained by using the hierarchical MLKSFM with 4 classes pre-classified in the lower level MLKSFM. The results indicate that with the same LID system configuration, the MFCC + MODGDF features outperform the MFCC only for most of the cases.

5. CONCLUSION AND DISCUSSION

The results in this paper have shown that the multi-layer Kohonen self-organizing feature map achieves promising performance for the OGI-TS LID task. By combining the MODGDF feature with MFCC and using a hierarchical structure of the MLKSFM, the LID rate can be significantly increased. The use of combined MFCC and MODGDF features provides a better LID rate compared with the use of MFCC only or MODGDF only feature, for most of the cases. The best LID rate for the 10-sec speech utterances is achieved by using MLKSFM and for the 45-sec speech utterances, the best LID rate is obtained by using hierarchical MLKSFM with 4 classes pre-classified on the lower level MLKSFM.

Our future work will concentrate on combining some higher level language features in the input feature vectors.

6. REFERENCES

[1] Zissman, M., "Comparison for Four Approaches to Automatic Language Identification of Telephone Speech", in *IEEE Trans. Speech and Audio Proc.*, vol. 4, pp. 31-44, 1996.

[2] Wang, L., Ambikairajah, E., and Choi, Eric, H.C., "Multi-lingual Phoneme Recognition and Language Identification Using Phonotactic Information", in *Proc. ICPR 2006*, vol. 4, pp. 245-248, 2006.

[3] Singer, E., Torres-Carrasquillo, P. A., Cleason, T. P., Campbell, W. M., and Raynolds, D. A., "Acoustic, Phonetic and Discriminative Approaches to Automatic Language Recognition", in *Eurospeech in Geneva*, ISCA, pp. 1345-1348, 2003.

[4] Kohonen, T., "Self-Organizing Maps", in *Springer Series in Information Sciences*, vol. 30, 1995.

[5] Kohonen, T., Barna, T., and Chrisley, R., "Statistical Pattern Recognition with Neural Networks: Benchmarking Studies", in *Proc. ICNN*, vol. 1, pp. 61-68, July, 1988.

[6] Kohonen, T., Self-Organization and Associative Memory (2nd edition), Springer-Verlag Publishers, 1988.

[7] Wang, L., Ambikairajah, E., and Choi, Eric, H.C., "Multi-Layer Kohonen Self-Organizing Feature Map for Language Identification", in *InterSpeech 2007, accepted and to appear*.

[8] Murthy, H., and Gadde, V., "The Modified Group Delay Function and Its Application to Phoneme Recognition", in *Proc. ICASSP 2003*, vol. 1, pp. 68-71, 2003

[9] Wang, L., Ambikairajah, E., and Choi, Eric, H.C., "Automatic Language Recognition with Tonal and Non-Tonal Language Pre-Classification", in *15th European Signal Processing Conference, 2007, accepted and to appear*.

[10] de la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C., and Rubio, A. J., "Histogram Equalization of Speech Representation for Robust Speech Recognition", in *IEEE Trans. Speech and Audio Proc.*, vol. 13, issue 3, pp. 355-366, 2005.

[11] Kohonen, T., Kaski, S., Lagus, K., and Honkela, T., "Very Large Two-Level SOM for the Browsing of Newsgroups", in *ICANN 1996*, LNCS, Springer Berlin, vol. 1112, pp. 269-274, 1996

[12] Tomczyk, A., Szczepaniak, P. S., and Lis, B., "Generalized Multi-Layer Kohonen Network and Its Application to Texture Recognition", in *ICAISC 2004*, LNCS, Springer Berlin, vol. 3070, pp. 760-767, 2004.

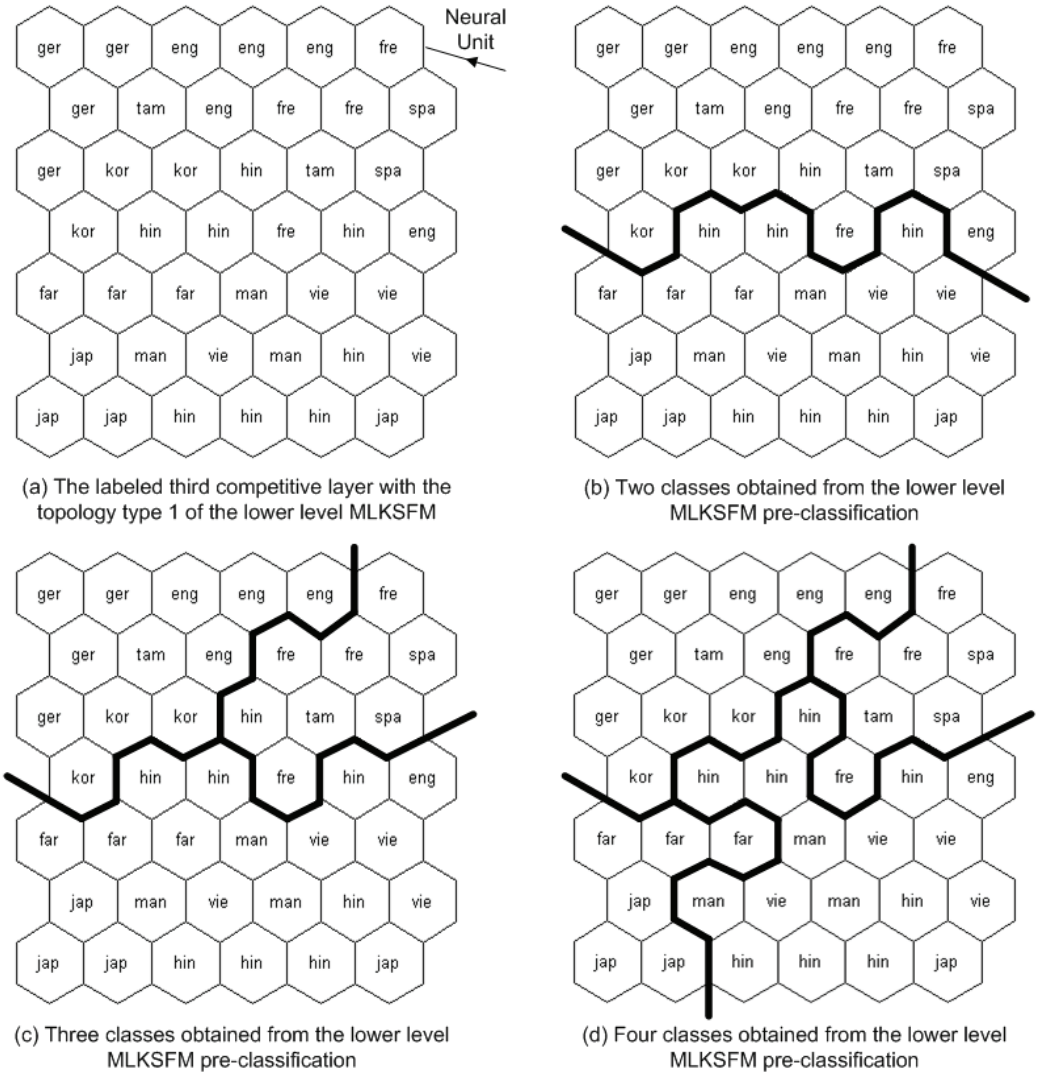

Fig. 6. The labeled third competitive layer in the MLKSFM

TABLE 2. COMPARISON OF LID RATES FOR MLKSFM AND HIERARCHICAL MLKSFM SYSTEMS

Number of language classes in the lower level MLKSFM		N.A.	2	3	4	11
Number of languages handled by each of the upper level MLKSFM		11	5&6	3&3&5	3&3&3&2	N.A
LID rate (%) on 45-sec speech	With MFCC only	78.2	74.5	70.0	80.9	63.6
	With MODGDF only	76.4	73.6	70.9	78.2	60.0
	With MFCC and MODGDF	83.6	81.8	78.2	**87.3**	65.5
LID rate (%) on 10-sec speech	With MFCC only	58.2	54.5	52.7	56.4	50.9
	With MODGDF only	55.5	50.0	50.9	52.7	47.3
	With MFCC and MODGDF	**60.0**	57.3	54.5	59.1	53.6