# Multi-Layer Kohonen Self-Organizing Feature Map for Language Identification

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

[1] School of EE&Telecom, the University of New South Wales, NSW, 2052 Australia
[2] ATP Research Laboratory, National ICT Australia, NSW, 1435 Australia
l.wang@student.unsw.edu.au, ambi@ee.unsw.edu.au, eric.choi@nicta.com.au

10.21437/Interspeech.2007-73

## Abstract

In this paper we describe a novel use of a multi-layer Kohonen self-organizing feature map (MLKSFM) for spoken language identification (LID). A normalized, segment-based input feature vector is used in order to maintain the temporal information of speech signal. The LID is performed by using different system configurations of the MLKSFM. Compared with a baseline PPRLM system, our novel system is capable of achieving a similar identification rate, but requires less training time and no phone labeling of training data. The MLKSFM with the sheet-shaped map and the hexagonal-lattice neighborhoods relationship is found to give the best performance for the LID task, and this system is able to achieve a LID rate of 76.4% and 62.4% for the 45-sec and 10-sec OGI speech utterances, respectively.

**Index Terms**: language identification, multi-layer Kohonen self-organizing feature map, histogram equalization

## 1. Introduction

Recent studies on language identification (LID) have explored a variety of methods that utilize different levels of speech features, including acoustics, phonotactic and prosodic features [1][2][3][4]. Extensive studies have been done based upon high-level phonotactic features of language identification. The best known method for this is PPRLM (Parallel Phone Recognition and Language Modeling) which has been shown to be quite successful [1][2]. In PPRLM, the input speech utterances are firstly mapped into phone sequences by a set of phone recognizers. Then the n-gram language model is employed to estimate the probability of the occurrence of a particular phone sequence for the final scoring. The phonotactic features are believed to carry more discriminative information about the language than acoustic features and prosodic features, and are believed to be more robust. The extraction of the phonotactic information, however, requires the speech to be labeled at a fine phone level for model training. This is a very time-consuming and expensive task, thus fine phone labeling is only available for a few languages. Lately, more research has focused on the Gaussian Mixture Model and Universal Background Model (GMM-UBM) based systems [3][4]. No phone labeled data is required for training the GMM-UBM based systems, however adaptation for the UBM requires considerable computational resources. Thus to design a language identification system without phonetic labeled training data that uses few computational resources is a challenging problem.

For a system to yield high performance in language identification, two important properties must be comprised: the ability to realize complex decision regions in the feature vector space, and the ability to extract sufficient information for the speech signal [1][3]. In this paper we propose a novel language identification system by using a segment-based Kohonen Self-Organizing Feature Map (KSFM) [5][6][7]. A KSFM 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 [5][6]. We will show that by using a segment-based input feature vector, the novel 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.

## 2. Multi-layer Kohonen self-organizing feature map for language identification

In our proposed system, the language identification task is regarded as a feature vector quantization problem. It is well known that the KSFM performs very well in vector quantization. However, we will show that the traditional KSFM is not adequate for the language identification task, and we will extend the single-layer KSFM to a multi-layer KSFM.

### 2.1. Single-layer KSFM for language identification

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

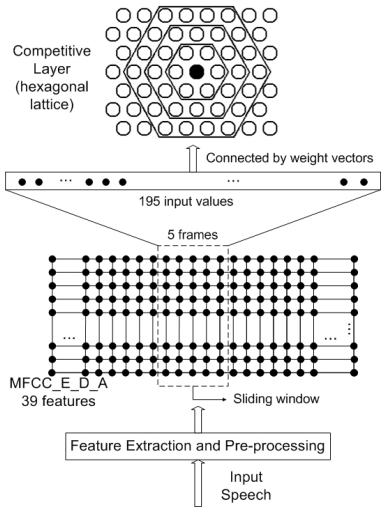


Figure 1: *The network structure of segment-based KSFM for language identification.*

Firstly, for each input speech utterance, the basic feature vector is extracted. The basic feature 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. Feature normalization is used in pre-processing. Normalizing the variables is important so that no single value has an overwhelming influence on the training result. In this experiment we use the histogram equalization (HEQ) as the normalization. The HEQ maps the histogram of each component of the feature vector onto a reference histogram [8]. The HEQ is widely used in speech recognition also because its ability to compensate for the effect of noise processes distorting the feature space [8]. In order to keep temporal information in the speech signal, we define a 5-frame window which moves over the whole speech utterance with a shift of one frame between windows. Each window position yields a training vector of 195 dimensions (5 frames * 39 MFCC). Each segment-based training feature vector 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 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 $\boldsymbol{x}$ is found at neuron $\boldsymbol{C}$, then we have

$$\| x - \underline{w}_C \| = \min_j \| x - \underline{w}_j \| \quad (1)$$

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

If $\boldsymbol{n}$ is used to denote a discrete time index, then the weight vector is updated according to

$$\underline{w}_j(n+1) = \begin{cases} \underline{w}_j(n) + a(n)(x(n) - \underline{w}_j(n)), & for\ j \in N_C \\ \underline{w}_j(n), & otherwise \end{cases} \quad (2)$$

where $\boldsymbol{a(n)}$ is a positive constant that decays with time, and $\boldsymbol{N_C}$ defines a topological neighborhood around the best matched neuron unit $\boldsymbol{C}$, 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 $\underline{\boldsymbol{w}}_{\boldsymbol{j}}$ 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 $\boldsymbol{X}$, the set of segmented, normalized feature vectors are first calculated as $\boldsymbol{X = \{ x_1, x_2, ..., x_t, ..., x_T \}}$. For each feature vector $\boldsymbol{x_t}$ the best match is found from each of the neural units in the competitive layer with a corresponding weight vector $\underline{\boldsymbol{w}}_{\boldsymbol{j}}$. The label $\boldsymbol{i}$ in $\underline{\boldsymbol{w}}_{\boldsymbol{j}}$ is added to the corresponding feature vector $\boldsymbol{x_t}$, where $\boldsymbol{i \in \Lambda, \Lambda = \{ 1, 2, \ldots, M \}}$ and $\boldsymbol{M}$ is the number of target languages. The final identification is performed by using a voting function:

$$\Phi(X) = \arg\max N(i \mid X), \quad i \in \Lambda \quad (3)$$

where

$$N(i \mid X) = \sum_{t=1}^{T} \tau(x_t \in i) \quad (4)$$

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

$$\tau(x_t \in i) = \begin{cases} 1, \text{if the unit for language } i \text{ is the best matched} \\ 0, \text{otherwise} \end{cases} \quad (5)$$

### 2.2. Multi-layer KSFM

The traditional single-layer KSFM is capable of preserving the topological neighborhood relationship of the input vector space and faithfully describing the distribution of data points embedded in a high-dimensional space onto a plane. Sometimes however, in more complex tasks, it proves to be very complicated to design an appropriate competitive layer, in terms of number of neural units, the neighborhood relationship of neural units (hexagonal lattice, rectangular lattice, etc.) and the shape of the map (sheet, cylinder, toroid, etc.) [9][10]. This can be caused either by difficulty utilizing the existing feature vectors, or insufficient knowledge about the problem domain. In our language identification task, where non-labeled low-level acoustic features are used, a frame-based feature vector with very high dimensionality is employed in order to capture the temporal information. In such a situation the multi-layer KSFM (Fig. 2) can be of use.

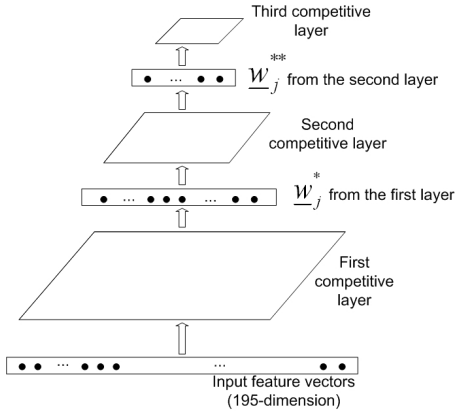


Figure 2: *The structure of multi-layer KSFM.*

The multi-layer KSFM (MLKSFM) used for language identification task is organized as a pyramidal structure consisting of multiple layers of single-layer KSFMs [9]. The number of neurons in a 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 (3).

## 3. Experiments

We used the PPRLM system as a baseline system. The OGI-TS speech corpus was used to perform the language identification task on the single-layer KSFM, multi-layer KSFM and the baseline system. 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.

### 3.1. System configuration

For the baseline PPRLM system, six phone recognizers were trained by using all the labeled data from OGI-TS, and the Witten-Bell discounting method was used in the language modeling in order to improve the LID rate [2].

For the single-layer KSFM system, the topology of the competitive layer was with the sheet shaped map and the hexagonal lattice neighborhoods relationship. The size of the competitive layer was defined as 50*30.

For the MLKSFM LID system, different types of topologies were compared in the competitive layer. For the local lattice structure, the hexagonal grid and the rectangular grid (Fig. 3) were used, while sheet and cylinder shapes (Fig. 4) were used to indicate the global map shape. 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.

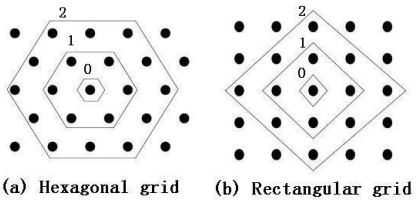


Figure 3: *Different lattice structures of the MLKSFM.*

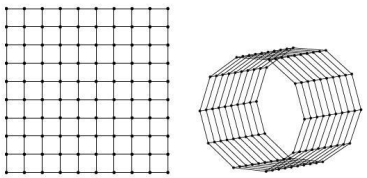


Figure 4: *Different map shapes of the MLKSFM.*

An additional evaluation was also performed by using the MLKSFM on the LID task. In the additional test, the basic feature only contained 12 Mel-frequency cepstral coefficients (MFCC) and the log-energy, which resulted in a 13-dimensional vector. The 5-frame windows was still used to generate a 65-dimensional (5 frames * 13 MFCC) feature vector. In the additional evaluation, we increased the map size of the first layer to 150*45. The sizes for the second and third layers were unchanged. The competitive layer was designed by using a sheet-shaped map with hexagonal grid.

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.

### 3.2. Experimental results

Table 1 shows the LID rates for different systems. 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 sheet shaped map and the hexagonal lattice neighborhoods relationship. The results indicate that the single-layer KSFM is not able to perform well for the language identification task. With the sheet shaped map topology (topology 1 and 2) in the competitive layer of the MLKSFM, we achieve significant improvement on 10-sec utterances and somewhat minor improvement on 45-sec utterances. The MLKSFM with the cylinder shaped map (topology 3 and 4) performs slightly worse than the baseline PPRLM system on both the 45-sec and 10-sec utterances. Our experiments also prove that the MLKSFM system requires less computation time than the PPRLM system in both the training and testing sessions. The training time for the baseline system is approximately 1.7 * real-time (i.e. a 10-sec utterance takes about 17-sec for processing), compared to 1.2 * real-time for averaged training time for different configurations of MLKSFM. For testing, the baseline system takes 0.39 * real-time compared to 0.13 * real-time for the MLKSFM.

Table 1. *Comparison of LID rate for different systems.*

| | | 45-sec | 10-sec |
|---|---|---|---|
| Baseline PPRLM | | 71.4% | 52.8% |
| Single-layer KSFM, with 195-dimensional input feature vector | | 51.9% | 41.2% |
| Multi-layer KSFM | **topology 1 (hexagonal grid with sheet shaped map), with 195-dimensional input feature vector** | **76.4%** | **62.4%** |
| | **topology 2 (rectangular grid with sheet shaped map), with 195-dimensional input feature vector** | 74.1% | 59.3% |
| | **topology 3 (hexagonal grid with cylinder shaped map), with 195-dimensional input feature vector** | 68.6% | 44.3% |
| | **topology 4 (rectangular grid with cylinder shaped map), with 195-dimensional input feature vector** | 69.3% | 42.0% |
| | **topology 1, with only 65-dimensional input feature vector** | 72.4% | 53.7% |

It is worth noting that for the MLKSFM with the reduced dimensionality of the input feature vector, the proposed LID system still outperforms the baseline PPRLM system. This is probably because some redundant information exists in the MFCC and its associated delta and acceleration coefficients. Additionally, increasing the map size in the first competitive layer can compensate for the effect of reducing the number of input feature vectors. It should be noted that by reducing the dimensionality of the input feature vector, the training and testing computation for the additional test is largely reduced.

Fig. 5 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.

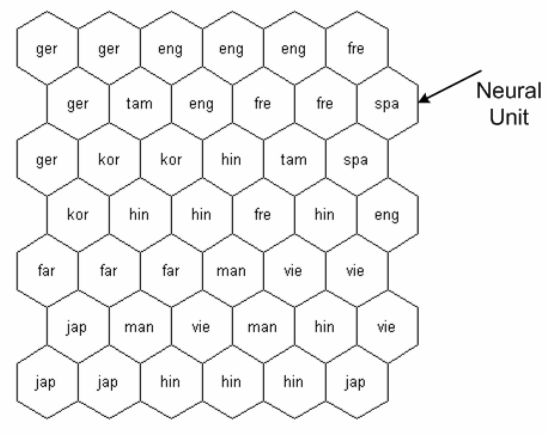


Figure 5: *The labeled third competitive layer with the topology type 1 in the MLKSFM.*

In Table 2 we report the experiment results on a pair-wise LID task by using the MLKSFM system with the hexagonal grid and sheet shaped map. The pair-wise LID rates of the baseline PPRLM system on 45-sec utterances are also shown in square brackets as a reference. Compared with the PPRLM system, the novel MLKSFM system performs better for almost all pair-wise identification tasks; only 9 out of 55 pairs perform worse.

## 4. Conclusion and discussion

The results in this paper have shown that the multi-layer Kohonen self-organizing feature map gives promising results in the LID task compared with the baseline PPRLM system. By using an appropriate mapping topology, the LID rate is significantly increased. The best LID rate is achieved on MLKSFM with the sheet-shaped map and the hexagonal lattice neighborhoods relationship. Also, compared with the baseline PPRLM system, the novel MLKSFM system requires less computation time and does not need any phone labeled training data.

Our future work will concentrate on removing redundant information in the input feature vector, thus to reduce the size of the feature parameters. Also the relationship of different topologies and the LID rate will be examined in detail. Additionally, we will explore methods of including some higher level language features in the input feature vectors.

## 5. 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] Lin, C. Y., and Wang, H. C., "Language Identification Using Pitch Contour Information in the Ergodic Markov Model", in *Proc. ICASSP 2006*, vol. 1, pp. 193-196, 2006.

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

[6] 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.

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

[8] 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.

[9] 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

[10] 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.


Table 2. *Confusion matrix of LID rate (%) in pair-wise LID task (the pair-wise LID rate of the baseline PPRLM system on 45-sec utterances is also listed in the square brackets).*

| 10/45-sec | Far | Fre | Ger | Hin | Jap | Kor | Man | Spa | Tam | Vie |
|---|---|---|---|---|---|---|---|---|---|---|
| Eng | 73.3/**86.7** [73.3] | 66.7/76.7 [80.0] | 66.7/**78.3** [76.7] | 63.3/**81.7** [80.0] | 76.7/**88.3** [83.3] | 70.0/**85.0** [83.3] | 78.3/**86.7** [80.0] | 61.7/**75.0** [70.0] | 71.7/**88.3** [83.3] | 69.3/**85.0** [80.0] |
| Far | / | 83.3/**91.7** [86.7] | 75.0/**88.3** [80.0] | 61.7/**78.3** [75.0] | 70.0/**81.7** [81.7] | 66.7/76.7 [81.7] | 60.0/75.0 [88.3] | 58.3/**75.0** [63.3] | 66.7/**75.0** [71.7] | 61.7/**78.3** [73.3] |
| Fre | / | / | 65.0/**80.0** [78.9] | 69.3/**83.3** [79.2] | 70.0/**85.0** [81.7] | 66.7/**81.7** [76.7] | 73.3/**80.0** [78.3] | 66.7/76.7 [80.0] | 71.7/**78.3** [73.3] | 66.7/75.0 [83.3] |
| Ger | / | / | / | 66.7/**83.3** [80.0] | 69.3/**85.0** [80.0] | 63.3/75.0 [81.7] | 69.3/**83.3** [76.7] | 73.3/**80.0** [78.3] | 65.0/**75.0** [73.3] | 78.3/**91.7** [68.3] |
| Hin | / | / | / | / | 63.3/78.3 [80.0] | 65.0/**76.7** [76.7] | 70.0/**80.0** [78.3] | 71.7/**83.3** [78.3] | 66.7/**80.0** [75.0] | 66.7/**78.3** [76.7] |
| Jap | / | / | / | / | / | 76.7/**85.0** [81.7] | 61.7/76.7 [85.0] | 69.3/**81.7** [73.3] | 73.3/**83.3** [78.3] | 61.7/**75.0** [71.7] |
| Kor | / | / | / | / | / | / | 68.3/**80.0** [78.3] | 70.0/**81.7** [73.3] | 68.3/**85.0** [83.3] | 73.3/**83.3** [76.7] |
| Man | / | / | / | / | / | / | / | 73.3/**83.3** [78.3] | 71.7/**86.7** [73.3] | 66.7/75.0 [78.3] |
| Spa | / | / | / | / | / | / | / | / | 66.7/**78.3** [75.0] | 69.3/**80.0** [66.7] |
| Tam | / | / | / | / | / | / | / | / | / | 71.7/**83.3** [73.3] |