Emergence of Language-Specific Phoneme Classifiers in Self-Organized Maps

Marek W. Doniec, Brian Scassellati, Willard L. Miranker

***Abstract*— The difference between self-organizing maps based phoneme classifiers that emerge for different input languages is studied. For each such language a self-organizing map is trained on Mel-Frequency Cepstral Coefficient (MFCC) converted auditory input to form a phoneme classifier. Unsupervised learning is used as the training method. The emerging classes are then compared to the classes found in the International Phonetic Alphabet. Particular class differences across languages and speakers are discussed. We show that SOMs adapt to speakers and languages, even when only given a small training data-set. Additionally, we show that some neurons in SOMs react only to input in one of the two trained languages and that some neurons can be used as word boundary classifiers.**

I. Introduction and Related Work

Kepuska et al. [1] have shown that a hexagonal lattice self-organizing map (SOM) shows similar response patterns for the same words and different response patterns for different words. They used 9 repetitions of 20 different words to train and test their SOM. Kumpf et al. [2] showed that using a Hidden Markov Model (HMM) they were able to classify accents within a group of Australian English speakers with an accuracy of up to 85.3%. Kangas [3] has shown that using a time-dependant representation of Mel-Frequency Cepstral Coefficients (MFCCs) can improve phoneme classification from a 10.4% rate error to a 5.0% rate error. However none of these works have compared the resulting classes to the classes found in the international phonetic alphabet (IPA). This alphabet is a much studied and widely accepted classification of phonemes that provides a representation for phonemes of any spoken language [4]. A comparison of the classes learned by a phoneme-recognition SOM to the IPA might reveal strengths or weaknesses of training phoneme classifiers using SOMs and possibly lead to improvement. Further a positive correspondence would suggest that SOMs are capable of capturing the functionality of the human auditory system.

We investigate the differences between phone classes of different languages. The languages are chosen to be different enough so that a native speaker of one language usually has a strong accent in the other language chosen. We first convert the audio signal using Mel Frequency Cepstral Coefficients (MFCCs) which approximate the human auditory system's response and are widely used in speech recognition systems [5], [6]. A self-organizing map is then trained on feature vectors for each of the languages tested. We use unsupervised learning. The classes found in the resulting feature maps are then compared to the IPA by submitting example words for specific phonemes to the trained SOMs or by looking at neurons that respond only to utterances from a particular language. In particular we look for classes that are present in at least one of the trained SOMs but not present in the other trained SOMs. In addition we trained an SOM on two languages and examined the neurons in this SOM that responded only to utterances from one of the two languages. We then identified the phoneme class that these neurons correspond to. Finally we investigate the use of Principal Component Analysis (PCA) to find phoneme classes and to compare phoneme classes from different languages.

The paper is organized as follows. Section II explains how data was collected, preprocessed, and how the SOMs were trained. Section III talks about differences between SOMs that were trained on utterances from different speakers and in different languages. Section IV focuses on differences between languages. Section V examines the use of PCA to detect language and speaker dependencies. Section VI summarizes the results and Section VII contains brief critique.

II. Methodology

First we describe the setup for recording our wave samples. Then we describe how the self-organizing maps were trained, and we introduce a distance measure for the trained SOMs.

A. Recording

Wave files for the experiment were recorded at 8 bits mono with a 22 kHz sampling rate. Subjects were presented with text excerpts from newspaper articles, encyclopedia entries, and stories. A simple computer microphone was used for recording and it was placed in front of the subject at a distance of about 50 cm. We recorded four speakers. The first three speakers are native German speakers and were recorded reading German and English texts. The fourth speaker is a native Polish speaker and was recorded reading Polish as well as English texts. The first speaker is male, the others are female. The first speaker is the first author of this paper. For each language / speaker, four wave files were recorded for a total of 32 wave files. Each wave file had a 120 second duration and is about two paragraphs of text long. In the following, the labels S_{1E} and S_{1G} denote the first speaker in English and German respectively. The labels S_{2E}, S_{2G}, S_{3E}, and S_{3G} denote the second and third speaker in English and German respectively. The fourth speaker is referred to by the labels S_{4E} and S_{4P} for English and Polish. $S_{1E,1}$ stands for the first wave file recorded for the first speaker in

Marek W. Doniec, Brian Scassellati, and Willard L. Miranker are with the Department of Computer Science, Yale University, New Haven, CT 06511, USA (email: marek.doniec@yale.edu, scaz@cs.yale.edu, miranker@cs.yale.edu).

	S_{1E}	S_{1G}	S_{2E}	S_{2G}	S_{3E}	S_{3G}	S_{4E}	S_{4P}
S_{1E}	11.8	15.0	62.1	61.5	61.3	51.5	48.9	48.3
S_{1G}	15.0	10.3	52.9	52.2	54.0	44.4	47.9	42.7
S_{2E}	62.1	52.9	12.3	16.7	26.1	28.6	32.1	24.2
S_{2G}	61.5	52.2	16.7	11.7	24.0	22.7	30.3	22.5
S_{3E}	61.3	54.0	26.1	24.0	12.7	20.2	30.3	27.9
S_{3G}	51.5	44.4	28.6	22.7	20.2	13.2	30.3	27.2
S_{4E}	48.9	47.9	32.1	30.3	30.3	30.3	10.9	19.1
S_{4P}	48.3	42.7	24.2	22.5	27.9	27.2	19.1	12.2

TABLE I

AVERAGE DISTANCES FOR SOMS THAT WERE TRAINED ON DATA FROM EIGHT GROUPS (FOUR SPEAKERS, EACH IN TWO LANGUAGES). FOR EACH SPEAKER/LANGUAGE GROUP FOUR SOMS WERE TRAINED, ONE ON EACH WAVE FILE IN THAT GROUP. THIS RESULTED IN A TOTAL OF 32 SOMS AND 1024 DISTANCES. DISTANCES WERE COMPUTED ACCORDING TO SECTION II-C. EACH ENTRY IN THIS TABLE IS THE AVERAGE OF ALL 16 DISTANCES COMPUTED FOR THAT PARTICULAR PAIRING. THE STANDARD DEVIATION FOR ALL REPORTED AVERAGES IS $< 10\%$.

	S_{1E}	S_{2E}	S_{2G}
S_{1E}	0.28	0.72	0.80
S_{2E}	0.72	0.21	0.37
S_{2G}	0.80	0.37	0.25

TABLE II

AVERAGE DISTANCES BETWEEN THE FIRST PRINCIPAL COMPONENTS OF THE INPUT DATA, SORTED BY GROUP (LANGUAGE, SPEAKER). THE ABSOLUTE VALUES ARE OF NO MEANING, HOWEVER THERE IS A SIGNIFICANT DIFFERENCE BETWEEN THE DISTANCE FOR SPEAKERS AND THE DISTANCE FOR LANGUAGES. THIS SUGGESTS THAT PCA IS ABLE TO CAPTURE SPEAKER DIFFERENCES.

English, etc. The index for wave files extends to the other abbreviations appropriately.

B. Training the Self-Organizing Maps

Recorded wave files were first processed using the Matlab MFCC library [7]. The library offers tools to convert the entire wave file into 20-dimensional MFCCs. The signal was converted using a Hamming window of size 32 msec and a hop time of 16 msec. For example, a 32 second wave file would result in 2000 MFCCs of size 20. This fine grained resolution (62.5 Hz) was chosen to account for fast changing phonemes like fricatives.

All the SOMs trained in this paper are of size 10×10 neurons. Training data sets were chosen from the above mentioned set of preprocessed wave files according to each experiments specification. During training the entire selected training data was repeatedly presented as an epoch to the SOM being trained (a total of 20 times). We used competitive learning with a Gaussian neighborhood function and a learning coefficient that decreased after each epoch.

C. A Distance Measure

For each SOM N and $i \in \{1, ..., 100\}$ let $N(i) \in \mathbb{R}^{20}$ be the weight vector of the i^{th} neuron of N. For $j \in \{1, ..., 20\}$ let $N(i, j)$ be the j^{th} entry of the i^{th} weight vector of N. An advantage of this notation is, that a 10×10 SOM N can be represented by a 100×20 matrix in which each row represents one neuron. Define the distance between two neurons to be the square of the euclidian distance of their weight vectors:

$$d'(N_1(a), N_2(b)) = \sum_{k=1}^{20} (N_1(a, k) - N_2(b, k))^2$$

Further define the bijective function $m : 1, ..., 20 \rightarrow 1, ..., 20$ to be the optimal match between the neurons of two SOMs using the metric just defined. This means that m minimizes the following function:

$$d(N_1, N_2) = \sum_{i=1}^{100} d'(N_1(i), N_2(m(i)))$$

Define this optimal match distance to be the distance between two SOMs. We see that this distance measure satisfies the three distance axioms:

1) $d(N_1, N_2) \geq 0$ and $d(N_1, N_2) = 0$ iff $N_1 = N_2$. (Obvious.)
2) $d(N_1, N_2) = d(N_2, N_1)$. (Obvious.)
3) $d(N_1, N_3) \leq d(N_1, N_2) + d(N_2, N_3)$. If m_{12} is the optimal match for N_1, N_2 and m_{23} is the optimal match for N_2, N_3 then the optimal match for N_1, N_3 is at least as good as $m_{23}(m_{12})$.

III. SPEAKER AND LANGUAGE DEPENDENCIES

In this section we examine the differences (distances) between SOMs that are trained on utterances from different speakers and in different languages. We computed the distances between SOMs trained separately for all 32 wave files ($S_{1E,i}$, $S_{1G,i}$, $S_{2E,i}$, $S_{2G,i}$, $S_{3E,i}$, $S_{3G,i}$, $S_{4E,i}$, and $S_{4P,i}$ $i \in 1, 2, 3, 4$). The process was repeated 5 times to obtain a good average. However it turned out that the SOMs converge so strongly that the differences across two SOMs trained for the same data-set are negligible and thus with a Matlab precision of 4 digits the distances computed for all 5 runs were the same. The results can be seen in Table I.

Note that the absolute value of each distance does not provide useful information, because it depends on the number of neurons used and the representation of the MFCCs. However since the number of neurons and the MFCC representation chosen are the same for all SOMs, comparing two distances is a relevant approach to seek meaning. First we notice that the distance for SOMs trained on the same speaker and the same language are closer to each other then all the

other SOMs. This means that the metric used does capture some difference between different speakers and languages. Note next that the distance between speakers seems to be larger than the distance between languages. This means that our SOMs characterize speaker dependencies more readily than language dependencies. However the SOMs are still capable of capturing the difference between languages for one speaker. Thus we decided to use only one speaker who spoke multiple languages in the balance of the experiments.

IV. Language Specific SOM Areas

In this section we attempted to measure and visualize differences between SOMs that were trained on utterances in multiple languages collected from one speaker. In order to be able to generalize we trained SOMs with data from all four subjects. However a single SOM was always trained with data from only one subject to avoid speaker dependencies. Instead of using our previously defined distance measure, we now use activation maps (specified in this section) to visualize similarities for different language inputs. Because we work only with single SOMs, the distance measure does not enter into this part of the study.

A. Training the SOM

A separate SOM was trained for each speaker. Only the first two out of four wave files from each language were used to train the SOMs for this experiment. For example, for speaker 1 we used $S_{1E,1}$, $S_{1E,2}$, $S_{1G,1}$ and $S_{1G,2}$ to train the SOM. Training is described in detail in Section II-B. Note that the rest of this section is based on the one particular SOM that resulted from such training (speaker 1). However this training process was repeated multiple times for all four speakers. While the resulting SOMs had a different spatial distribution of the neurons, they showed the same properties. These properties are presented in the following subsections.

B. Activation Maps

After training the SOM the remaining four wave files $S_{1E,3}$, $S_{1E,4}$, $S_{1G,3}$, and $S_{1G,4}$ were processed into MFCCs. The newly trained SOM was used to classify the input vectors. For each input vector the Euclidean distance to all neurons was computed. Each input vector was then assigned the number of the neuron with the smallest Euclidean distance to this input vector (That means this neuron fired for that particular input vector). A count was kept how often each neuron would respond to the input data stream. A separate counter was kept for each of the four data streams $\{English, German\} \times \{training, test\}$. The activation counts were then visualized in activation maps that are shown in Figure 1. In these maps the height of the bars in each neuron corresponds to the firing frequency for a given data stream. These streams are (from left to right) $\{S_{1E,1}$, $S_{1E,2}\}$, $\{S_{1E,3}$, $S_{1E,4}\}$, $\{S_{1G,1}$, $S_{1G,2}\}$, and $\{S_{1G,3}$, $S_{1G,4}\}$. English and German (Polish in case of speaker 4) are color coded red and green respectively. A neuron's shade intensity corresponds to

$$brightness = ||log(f_{red}/f_{green})||$$

where f_{red} and f_{green} are the firing frequencies of the two languages. The further apart the firing frequencies are, the higher is this value and the brighter the neuron is colored. That means that bright neurons responded predominantly to one of the two languages. The two activation maps shown have been trained and tested on utterances from speaker 1 (Figure 1(a)) and speaker 4 (Figure 1(b)). The activation maps for speaker 2 and speaker 3 are not shown because they show similar properties as Figures 1(a) and 1(b).

C. Observations

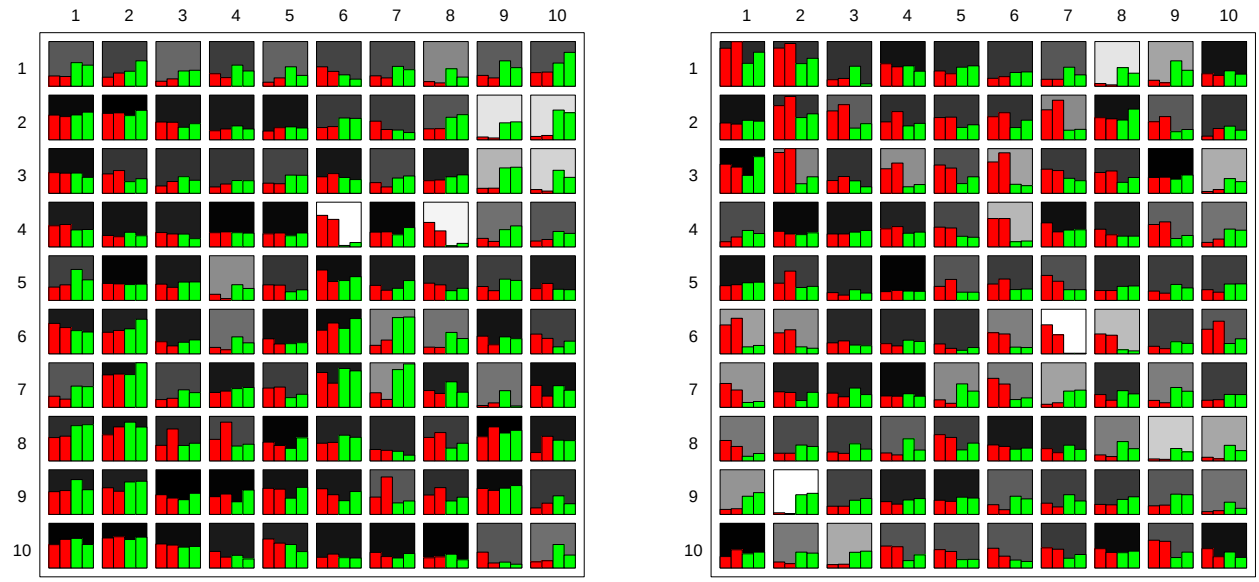
Note that the SOM develops the following four types of neurons:

1) There are a few neurons that respond sparsely to utterances from either language. For example, neuron $(10, 9)$ in Figure 1(a) and neuron $(9, 10)$ in Figure 1(b) show very low firing frequencies (For neuron numbering, see the caption of Figure 1). These neurons are most likely a result of the neighborhood-rule, i.e. two neighboring neurons that are far apart 'pull' this neuron into a space that is not used by the input.
2) Most neurons respond in similar ways to utterances in both languages. This is to be expected. Examples are neurons $(10, 1)$ and $(10, 2)$ in Figure 1(a). These neurons have a dark background in the figures.
3) Some neurons respond almost exclusively to utterances in German (or Polish in the case of speaker 4). One such neuron is neuron $(2, 10)$ in Figure 1(a). These neurons have a light background in the figures.
4) Some neurons respond almost exclusively to utterances in English. One such neuron is neuron $(4, 6)$ in Figure 1(a). These neurons also have a light background in the figures.

The occurrence of neurons that respond only to utterances in one language is a sign that the SOM does develop language-specific regions. The similarity in response patterns to training-set and test-set data shows that these regions are not entirely training-set dependant. As can be expected the response frequencies are not identical, however the predominance of certain similarities supports the hypothesis that SOMs develop language specific neurons. In the example illustrated in Figure 1(a) we can see that the upper right corner is German-dominated and that the center of the SOM is English-dominated. The left bottom corner responds frequently for both languages. As we shall see later it corresponds to silence (i.e., a pause) between words.

D. A Closer Look at Single Neurons

To illustrate the occurrence of each of the four neuronal classes discussed above and to show that language dependant neurons emerged during training, we have singled out the parts of the wave files that activate certain neurons that are predominant in one or in both languages, as the case may be. We give the total response time of neurons to utterances in different languages. The total response time for a neuron is calculated by multiplying the number of input vectors to



(a) Speaker 1: English (red) & German (green).

(b) Speaker 4: English (red) & Polish (green).

Fig. 1. These **SOM activation maps** display the response frequency of each of 10×10 neurons to utterances in both languages spoken by each speaker. For every neuron four bars represent the response frequency of that neuron to utterances from the following data sets (from left to right): [language 1, training set], [language 1, test set], [language 2, training set], [language 1, test set]. The taller a bar, the more a neuron responded to utterances from the associated data set. Languages are color coded (first language in green, second language in red). The background shade of each neuron corresponds to how language specific a neuron is. Dark shaded neurons fire equally often in response to data from both languages. Light shaded neurons fire predominantly in response to utterances in one of the two languages and fire only infrequently in response to utterance in the other language. For each SOM the neurons are numbered lexicographically in row/column order.

which this neuron responded by the stepping size that was used to calculate the input vectors (16 msec).

To see how well our SOM responded to single phonemes, we used it to classify recordings of single vowels spoken by speaker 1. A single and unique neuron responded to each utterance. The corresponding neurons are shown in Figure 2(d). This shows that the SOM is very well capable of distinguishing single phonemes.

Then we examined neuron number $(10, 1)$ which was very frequently activated for both data streams. We found that this neuron responded to MFCCs that represented silence. For the training set this neuron responded for a total of 7.6 sec for German and 5.75 sec for English. For the test set the total response time was 12.3 sec for German and 7.75 sec for English. This suggests that when speaking German, our speaker paused longer between words. However pauses might also have been classified by neighboring neurons. Since pauses occur frequently between words in both languages, the neurons corresponding to pauses were predominant in all activation maps. In Figure 1(b) for example, neuron $(1, 1)$ represented silence. To demonstrate this effect further Figure 2(a) shows neurons that responded to recorded silence. Although in the case of this particular SOM there are 8 neurons that respond to silence, they are all clustered together and still allow for an easy recognition of silence.

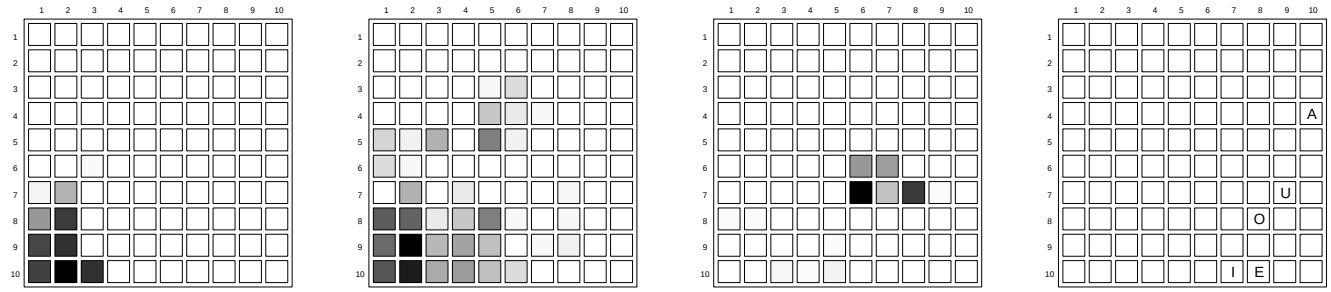
For further analysis we examined a neuron that exhibited a strong response only for MFCCs created from English utterances. Neuron $(4, 6)$ is predominantly red and responded a total of 0.05 sec for German and 3.15 sec for English in the training set and 0.05 sec for German and 3.15 sec for English in the test set. Here is a list of some of the words during which neuron $(4, 6)$ fired. Each word is accompanied by a pronunciation transcription as presented by the Merriam-Webster Online Dictionary [8].

- thirty [’th&r-tE]
- traverse [tr&-’v&rs]
- effort [’e-f&rt]
- computer [k&m-’py\u-t&r”]
- service [’s&r-v&s]
- aircraft [’er-\kraft”]

The neuron responded in particular to the [&r], which is pronounced like the ur/er in further.

Neuron $(7, 7)$ was also examined. It shows a much stronger activation for the German utterances. It responded for a total of 4.7 sec for German and 0.45 sec for English for the training set and 4.65 sec for German and 1.0 sec for English for the test set. This neuron corresponded to the nasal [n] sound as in ’nice’ which occurs less frequently in English then it does in German. Figure 2(c) shows neurons that responded to recordings of only the [n] sound spoken by speaker 1.

Similarly to neuron $(7, 7)$, neuron $(2, 10)$ responded more frequently to German than to English. It represented the [sh] sound as in ’shoe’. It responded for a total of 4.0 sec for German and 2.05 sec for English for the training set and 4.5 sec for German and 2.9 sec for English for the test set.



(a) Silence. (b) This & that. (c) N. (d) A, E, I, O, U.

Fig. 2. Activation maps for a bilingual SOM trained on utterances from speaker 1 (English and German). The SOM used to create these activation maps is also displayed in Figure 1(a). **a-c:** These activation maps show neuron response frequencies to several different utterances. The shade of the neuron corresponds to the firing frequency with a darker shade corresponding to a higher firing frequency. **d:** Neurons that respond to the vowels [a, e, i, o, u] are marked.

We also examined the [th] phoneme that does not occur in the German language but frequently occurs in English in words like 'this' and 'that'. A small test file that contained only the words 'this' and 'that' was recorded and the resulting MFCCs were classified with our bilingual SOM. The result is shown in figure 2. Multiple neurons responded to this input and we identify neuron $(5, 5)$ to be one of the neurons that responded to [th]. Neuron $(5, 5)$ responded for a total of 0.8 sec for German and 2.35 seconds for English for the training set and 1.45 sec for German and 3.6 sec for English for the test set. Neuron $(8, 5)$ also responded to [th] and showed similar total response times as neuron $(5, 5)$. While both neurons responded far more frequently to English then to German it is still surprising that these neurons responded to German at all, since the [th] sound does not occur in the German language. We see two possible reasons for this:

1) The most likely reason is that the speaker recorded was a native German speaker whose pronunciation very likely biased the result.
2) The resolution of the SOM might cause two sounds to be classified by the same neurons. Thus the same neurons might respond to similar sounds like [v].

V. Principal Component Analysis

We also investigated the use of Principal Component Analysis (PCA) to recognize speaker or language dependencies. PCA is a good candidate because it both extracts the most significant components and allows for a dimensionality reduction of the data. We hoped that we could identify components that would help identify the speaker or the language.

We used the files ($S_{1E,i}$, $S_{2E,i}$, $S_{2G,i}$, $i \in 1, 2, 3, 4$). We applied PCA to each of the MFCC data streams and saved the principal components (20 components for each file, each component of size 20). For each principal component and each pair of files we computed the Euclidian distance giving a total of $12 \times 12 = 144$ distances. The distances were then averaged over comparisons between files from the same group (there are three groups: S_{1E}, S_{2E}, and S_{2G}). This resulted in 20 tables, one for each principal component. Each table gives the average distances for this particular principal component between languages and speakers. We found that the first component represented the distances between speakers well as can be seen in Table II. However none of the components seemed to represent differences in language.

In a second analysis we used PCA to reduce the input space for our SOM training algorithm. For this the MFCC transformed files $S_{1E,1}, S_{1G,1}, S_{2E,1}$, and $S_{2G,1}$ were concatenated and PCA was applied to the resulting data stream. This time the representation in principal components was input to the SOM as input. The results were similar to those in Section IV.

However we found the following problem in utilizing PCA together with SOMs. PCA generates different principal components for different input files as shown in Table II. This results in the problem that if we transform two files separately then their representation in their respective principal components are of no value to the SOM, because they are independent of each other. We tried to represent additional data in the principal component space of the training data but only with marginal results.

VI. Summary of Results

We demonstrated that training SOMs on MFCCs results in SOMs that are both, speaker and language dependant. This result was obtained by comparing the SOMs with a specified metric. This suggests that SOMs can be used to differentiate between languages and between speakers.

We further demonstrated that SOMs do capture differences between languages that can be easily made discernable. In particular we demonstrated that if an SOM is trained with two languages then some neurons represent sounds that are unique or predominant in one of the languages. An additional result was that independently of the language used, the SOM has certain neurons that correspond to silence (a pause) and are activated more frequently than other neurons in the same SOM.

We demonstrated that PCA might be able to extract speaker differences but most likely is not suitable to extract language differences. Further we explained that it is

not possible to use PCA to preprocess the input to the SOM training algorithm because PCA will generate different principal components for two separate input samples. Thus representing each sample in the principal component space does not allow for a good comparison of two different samples.

VII. Discussion

The results suggest that training an SOM on unlabeled speech data can result in the formation of a phoneme classifier in which groups of neurons or single neurons correspond to different phonemes. Although these phonemes are not labeled they seem to represent the phoneme space of spoken languages well and correspond to phoneme classes found in the International Phonetic Alphabet. These SOMs help to further reduce the dimensionality of the input and could be of use for further classification and speech recognition tasks. The advantage over existing work is that our system uses unsupervised learning and thus needs no feedback.

We have found that the SOMs capture speaker as well as language differences and can adapt to speakers on a relatively small data set. This suggests that an SOM might be used for speaker identification or to adjust speech recognition systems to particular speakers. An analog of this is the preferential response of infants to their mother's voice as found by Mehler et al. [10]. In addition SOMs might show useful in developmental systems that first learn to discern phonemes from one speaker and then gradually develop speaker independence.

The language differences captured suggest that an SOM adapts to a certain language and its phonemes. This is also similar to findings in infants who habituate themselves to a particular set of phonemes and tend to attenuate non-native phonemes during advanced language learning [9]. Further we have shown that if trained with two languages at once an SOM can learn both phoneme sets and even distinguish between sounds that occur only in one of the languages. For future work we envision a system that learns to differentiate between several different languages based on the firing pattern of a trained SOM.

Another utilization of such an SOM could be speech segmentation. We observed that the neurons representing silence in the SOM fire more frequently then any other neuron (see Figure 1). This suggests that if indeed silence is the predominant feature vector, our system has developed a set of neurons that represent word-boundary signals. So not only does our SOM learn to classify phonemes but it could provide a subsequent speech recognition system with word boundary information.

We demonstrated that PCA is not well suited for preprocessing the input for the SOM in the case of building phoneme classifiers. We believe however that PCA might serve to extract principal components from a large data set for the identification of speakers. The reason that PCA demonstrated no utility for preprocessing is that different input files produced different principal components. This happens especially if the input files are small and thus provide only a small sample of training data. Further study will show whether the principal components will tend to stabilize for large bodies of data (multiple hours of recordings for each speaker / language combination). If such fixed points exist they might prove useful for preprocessing the speech signal.

Future work should include a more detailed analysis of the exhibited behaviors. The results obtained in this study are based on a relatively small data-set. We believe that a large scale study employing data from many subjects might reveal additional features and allow testing of the interaction between different languages and speakers. The reason that we used only one speaker for each SOM in the second part of the study is that currently there is no good method for extracting speaker independent feature vectors from speech. MFCC still captures the base frequency and possibly other speaker dependent features and thus does not allow for efficient comparison of languages across speakers. Current work on speaker independent phoneme classification usually trains classifiers on a large body of subjects [11]. While these classifiers learn to generalize across different subjects, they are still presented with speaker-dependent input such as MFCCs. A similar approach might be used in combination with the methods presented here. Naturally this would require a large body of data.

We have shown that the unsupervised learning of SOMs with as few as 100 neurons enables extraction of speaker and language differences. We showed that some can extract additional useful information such as word boundaries. Thus SOMs are well suited for use in unsupervised learning systems for word grounding (learning the meaning of words) and language recognition. We have also shown that SOM phoneme recognizers show learning and recognition behavior similar to that of human infants.

References

[1] V. Z. Kepuska and J. N. Gowdy, *Investigation of phonemic context in speech using self-organizing feature maps*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '89, Glasgow, Scotland, May 1989.

[2] K. Kumpf and R. W. King, *Automatic accent classification of foreign accented australian english speech*, In Proceedings of 4th International Conference on Spoken Language Processing, Philadelphia, USA, October 1996.

[3] J. Kangas, *Phoneme recognition using time-dependent versions of self-organizing maps*, In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, p. 101-104, Toronto, Canada, May 1991.

[4] International Phonetic Association, *Handbook of the International Phonetic Association - A Guide to the Use of the International Phonetic Alphabet*, July 1999.

[5] Z. Fang, Z. Guoliang and S. Zhanjiang, *Comparison of different implementations of MFCC*, Journal of Computer Science and Technology, 16(6), 582-589, 2001.

[6] *Mel frequency cepstral coefficient*, Wikipedia, http://en.wikipedia.org/wiki/Mel_frequency_cepstral_coefficient, 2006.

[7] *PLP and RASTA (and MFCC, and inversion) in Matlab*, D. Ellis, http://labrosa.ee.columbia.edu/matlab/rastamat/, 2006.

[8] *Merriam-Webster's Online Dictionary*, www.m-w.com, 2006.

[9] D. Burnham, *Language specificity in the development of auditory-visual speech perception*, R. Campbell & B. Dodd (Eds.), Hearing bye eye II: Adances in psychology of speechreading and auditory-visual speech. Hove, england: erlbaum UK, pp. 27-60, 1998.

[10] J. Mehler, J. Bertonicini, M. Barriere, *Infant recognition of mother's voice*, Journal of Perception, 7(5):491-7, 1978.

[11] M. Antal, *Speaker Independent Phoneme Classification in Continuous Speech*, Studia Univ. Babes-Bolyai Informatica, Vol. XLIX, No. 2, pp. 55-64, 2004.