ON PHONEMES AS COGNITIVE COMPONENTS OF SPEECH

Ling Feng, Lars Kai Hansen

Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads, B 321

ABSTRACT

COgnitive Component Analysis (COCA) defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, has been explored on phoneme data. Statistical regularities have been revealed at multiple time scales. The basic features are 25-dimensional short time ($20ms$) mel-frequency weighted cepstral coefficients. Features are integrated by means of stacking to obtain features at longer time scales. Energy based sparsification is carried out to achieve sparse representations. Our hypothesis is ecological: we assume that features that essentially independent in a context defined ensemble can be efficiently coded using a sparse independent component representation. This means that supervised and unsupervised learning should result in similar representations. We indeed find that supervised and unsupervised learning seem to identify similar representations, here, measured by the classification similarity.

Index Terms— Cognitive Component Analysis, Unsupervised Learning, Supervised Learning, Phoneme Classification.

1. INTRODUCTION

Cognition generally refers to capabilities of human minds, such as reasoning, perception, intelligence and learning, etc. The human cognitive system can model complex multi-agent scenery, and uses a broad spectrum of cues for analyzing perceptual input and for identification of individual signal process components. The purpose is to infer the proper action for a given situation. Robust statistical regularities can be exploited by an evolutionary optimized brain in making inference about appropriate actions [1]. *Statistical independence* is likely to be such regularity. Knowledge about an independence rule will allow the system to take advantage of a corresponding factorial code typically of (much) lower complexity than the one pertinent to the full joint distribution. The optimized representations of the low level cognition (perception) are known to be based on independence in the relevant natural ensemble statistics [2, 3]. This has led to a surge of interest in independent component analysis (ICA) for modeling perceptive tasks, and the resulting representations share many features with those found in natural perceptual systems. Examples are, e.g., in visual features [2, 3], and sound features [4].

Within an attempt to generalize these findings to a higher cognitive function, we proposed the cognitive component hypothesis which basically runs: *Human cognition uses information theoretically optimal ICA representations for generic data analysis*. COgnitive Component Analysis (COCA) is wherefore defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, see e.g., [5, 6]. In sensory coding it is proposed that visual system is near optimal in representing natural scenes by invoking 'sparse distributed' coding [7]. The sparse signal consists of relatively few large magnitude samples in a background of numbers of small signals. We envision that auditory areas of the perceptual system also abide by the sparse coding rule. When mixing such independent sparse signals in a simple linear mixing process, we obtain the 'ray structure' emblematic for cognitive component analysis. If a signal representation exists with a ray structure, ICA can be used to recover both the line directions (mixing coefficients) and the original independent source signals. Figure 1 illustrates the ray-structure representation of phoneme classification within three classes: vowels, fricatives, and stops.

Thus far, ICA has been used to model the ray structure and to represent semantic structure in text, social networks, and other abstract data, e.g. music [5, 8] and speech [9].

Since the mechanisms of human cognitive activity are still not fully understood, to quantify cognition may seem ambiguous and may also be considered way too ambitious. However, the direct consequence of cognition, human behavior, has a rich phenomenology that can be accessed and modeled. In the following analysis, we represent human cognition simply by a classification rule, i.e. based on a set of manually obtained labels we train a classifier using supervised learning. The question is then reduced to looking for similarity between the representations in supervised learning (of human labels) and unsupervised learning that simply explores the statistical properties of the domain. If a high correlation exists between the representations resulting from unsupervised and supervised

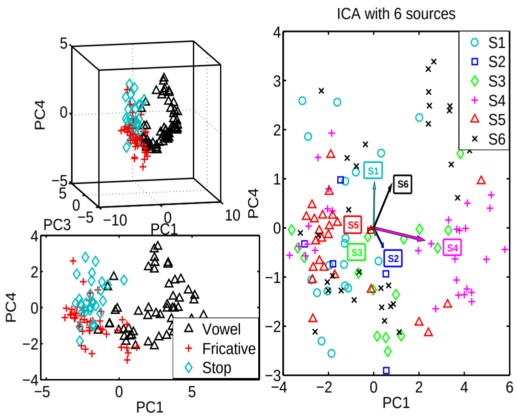

Fig. 1. Phoneme ray-structure. Figures on the left-hand side are scatter plots of phoneme features in the space of principal components. Data are displayed in different shapes denoting three classes: Vowels, Fricatives and Stops. Loosely speaking, fricatives and stops locate along solo-ray; and vowels spread more widely and can be represented by multi-rays. The right-hand side figure gives 6 independent sources. Arrows show the column vectors of the mixing matrix. By majority voting, source $1, 2$ stand for fricatives; $3, 4, 6$ for vowels; 5 for stops.

learning, we interpret this as the evidence that human cognition is based on the given statistical regularity. In this paper we will present a detailed comparison between unsupervised and supervised learning representations: at the classification rate level; at the sample-to-sample basis; and at the more detailed sample-to-sample posterior probability level. This paper focuses on cognitive component analysis of short time speech signals, to test whether phonemes are such cognitive components. First we discuss the preprocessing pipeline of COCA; secondly we introduce the unsupervised and supervised learning models; thirdly we systematically investigate the performance of unsupervised and supervised learning on the potential cognitive indicator: phoneme, and test whether the task is learnt in equivalent representations; and the conclusion summarizes this paper.

2. PREPROCESSING OF COGNITIVE COMPONENT ANALYSIS

Here we are going to elaborate on the speech-relevant cognitive component analysis. The basic preprocessing pipeline for speech COCA is shown in figure 2.

A efficient way of representing speech for machine speech analysis is usually to use spectral features of fairly low dimensionality, e.g. $20 \sim 30$ dimensions. The ideal features will be the ones which are capable of accounting for the functionality of human hearing system. The basic features in COCA analysis are extracted from digital speech signals leading to a fundamental representation that shares two basic aspects with the human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that the frequency resolution is better at lower frequencies. These so-called mel-frequency cepstral coefficients (MFCCs) can loosely represent the human auditory system response, which is triggered by the **mechanoreceptors** [10] of the inner ear, except that MFCCs can not model the outer ear which is critical for sound localization and loudness accuracy. The vibrations caused by the sound pressure waves receiving at the outer ear deflect the hairlike cells in the inner ear and trigger nerve impulses.

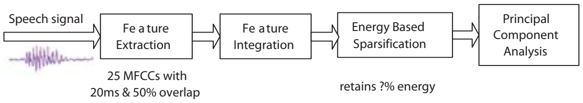
Speech signal
Fe a ture Extraction
Fe a ture Integration
Energy Based Sparsification
Principal Component Analysis
25 MFCCs with 20ms & 50% overlap
retains ?% energy

Fig. 2. Preprocessing pipeline for speech COCA. MFCCs are extracted at the basic time scale ($20ms$). Depending on the application features are integrated into longer time scales. Energy based sparsification is applied as a method to reduce intrinsic noise and get sparse representations. PCA projects features onto a base of cognitive processes. A subsequent ICA can be used to identify the actual ray coordinates and source signals.

The computation of MFCCs is based on the time-frequency analysis. Since speech signals are non-stationary, features must be extracted from short time intervals, i.e. $10 \sim 40$ ms. The fast fourier transform (FFT) transforms the convolution relation between the excitation sequence and the vocal system impulse response into production; and the logarithm, afterwards, provides us with the linear combination (addition between these two). The mel-frequency warping step changes the frequency scale from linear to mel-scale, which is approximately linear below $1k$Hz and logarithmic above. Finally discrete cosine transform (DCT) brings us to the mel-cepstrum. For detailed description, see [11].

2.1. Feature Stacking

For a feature to reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary, usually from one second to several tens seconds [12]. Feature integration is by and large a way to combine the information from several short-time features into a long-term feature. A simple integration is the stacking, in other words, vector 'concatenation' of signals.

1. Truncate speech signals into short time frames, $20ms$ long with 50% overlap;
2. Apply hamming window on each frame;
3. Extract MFCCs from each frame, which forms (e.g.) a 25-dimensional vector;
4. According to the time scale, the MFCCs from the first N frames are stacked into one $25 * N$-d vector;

5. Repeat 4 with the next N short time frames (without overlap) until all the short time frames are stacked (and exclude the residual).

The resulting $25 * N$-d features representing long time scales are then further processed.

2.2. Energy Based Sparsification (EBS)

Simple energy based filtering leads to sparse representations. Sparsification is regarded as a simple means to filter out the small signals, which emulates a saliency based *attention* process related to **detectability** and **sensory magnitude** from perceptual principles [10]. For auditory perception only the signals reaching the postsynaptic cell's threshold will lead to the cell firing [13]. Therefore sparsification is done by thresholding the stacked features, and only coefficients with superior energy are retained, and the rest is set zero.

2.3. Principal Component Analysis

PCA is an orthogonal linear transformation technique. It is often used for dimensionality reduction, and in the meanwhile remains the most variance of the data. In textual information analysis PCA is known as LSA. It presumes that the semantic content of the overall document can be approximated as the word usage. The low-dimensional space transformed by PCA/LSA from high-dimensional space is regarded as the basis for all cognitive processing [14]. LSA has human-like performance in text analysis, we assume that it can also be used to get the relevant basis for speech cognitive related tasks. It has been proved that in some cases, LSA can provide good simulations of human cognitive processes alone, and in other cases it is often operated as base for cognitive processes.

Singular value decomposition (SVD) is invoked to identify a relevant signal subspace based simply on signal variance,

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T, \;\; \mathbf{Y} = \mathbf{U}_k^T\mathbf{X}, \tag{1}$$

where **X** is a m-by-n data matrix, **U** is a m-by-m orthonormal matrix, Λ is a m-by-n matrix with the singular values along the diagonal, and **V** is a n-by-n orthonormal matrix. The dimensionality of data is reduced by projecting the data to the first k principal components ($k < m$).

3. MODELS

Having the comparison of the unsupervised and supervised learning in mind, we need to have two models which share similarities w.r.t the model structure. Moreover both models should allow sparse linear ray-like features. The Bayesian classifier which assumes a known probabilistic density distribution for each class, has been widely used and is misclassification error rate optimal. Here we choose two Bayesian classifiers: Naive Bayes and Mixture of Gaussians (MoG). For the unsupervised learning model we first apply unsupervised ICA only on the features. After recovering the source signals, we add the label information to a naive Bayes classifier, which assumes that the distribution of the source within each class is Gaussian. To keep the consistency of using Bayesian classifier and Gaussian model, we choose Mixture of Gaussians as the supervised learning model. This is a simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'?

3.1. Unsupervised Learning

As mentioned, if the sparse features are essentially independent, ICA can be used to recover both the mixing coefficients and the original independent sources. The typical algorithms for ICA use centering, whitening and dimensionality reduction as preprocessing steps in order to reduce the complexity of the algorithm. PCA is normally used to achieve the whitening and dimension reduction. Since in the preprocessing pipeline we have applied PCA on stacked and sparsified MFCC features, we directly apply ICA algorithm on PCA coefficients without dimensionality reduction.

The generative formula of noise free ICA model is

$$\mathbf{Y} = \mathbf{AS}, \tag{2}$$

where **Y** is the k-dimensional observation; **A** is the mixing matrix with dimension k-by-p; **S** is the matrix of p independent sources which are assumed non-Gaussian. ICA aims at estimating both the mixing matrix **A** and the sources **S**. This is done by either maximizing the non-Gaussianity of the calculated sources or minimizing the mutual information.

The original sources can be recovered by

$$\mathbf{S} = \mathbf{WY}, \tag{3}$$

where we assume the total no. of sources (k) is the same as the dimension of the observation **y** (p) in the following experiments, hereby $\mathbf{W} = \mathbf{A}^{-1}$ is the unmixing matrix, and the **A** and **W** matrices are therefore square.

To reveal the performance of unsupervised learning in classification tasks, we first train the unsupervised model using only the features (principal components) $\mathbf{Y}$ to recover the sources **S**. Since sources are independent, then naive Bayes classifier can be applied on sources with the training set labels. This is also referred to as unsupervised -then- supervised learning scheme.

The naive Bayes classifier assumes independency of input feature for each class, and is based on Bayes' theorem:

$$p(\mathbf{C}_i|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)} \tag{4}$$

where $p(\mathbf{C}_i)$ denotes the i^{th} class prior; $p(\mathbf{s}|\mathbf{C}_i)$ is the likelihood of the $\mathbf{C}_i$; and $p(\mathbf{C}_i|\mathbf{s})$ is the posterior of the i^{th} class given data **s**: $\mathbf{s} = (s_1, \ldots, s_p)^T$.

As naive Bayes assumes that the data input variables are independent, the likelihood in equation (4) can be simplified as:

$$p(\mathbf{s}|\mathbf{C}_i) = \prod_{n=1}^{p} p(s_n|\mathbf{C}_i), \tag{5}$$

where each $p(s_n|\mathbf{C}_i)$ is modeled as univariate Gaussian distribution $\mathcal{N}(\mu_{ni}, \sigma_{ni}^2)$.

For the classification problem, we apply the **W** learnt from training set to new data $\mathbf{Y}^{new}$, and recover their sources $\mathbf{S}^{new}$. Afterwards, the trained naive Bayes classifier with a set of Gaussian parameters (means and variances) will be used on $\mathbf{S}^{new}$ to predict the labels of new data.

3.2. Supervised Learning

As for the supervised learning model, we intend to choose a very flexible model, which is able to represent human decisions. We here use the Mixture of Gaussians,

$$p(\mathbf{C}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}, \tag{6}$$

and the likelihood will be,

$$p(\mathbf{y}|\mathbf{C}_i) = \sum_j p(\mathbf{y}|j, \mathbf{C}_i)p(j|\mathbf{C}_i), \tag{7}$$

where $p(\mathbf{y}|j, \mathbf{C}_i) = \mathcal{N}(\mathbf{y}|\mathbf{m}_{ji}, \mathbf{V}_{ji})$, and $p(j|\mathbf{C}_i)$ is the mixing parameters in class $\mathbf{C}_i$. The parameters $\mathbf{m}_{ji}$, $\mathbf{V}_{ji}$ are estimated from the training set via the standard Expectation-Maximization algorithm. For simplicity, we assume the covariance matrices to be diagonal. Note that although features are independent within each mixture component due to the diagonal covariance matrix, the mixture model does not factorize over features. The MoG is capable of modeling arbitrary dependency structures among features [15] if the number of mixture components is sufficiently large. On the other hand, a MoG with many mixture components is prone to overfitting and will most likely not generalize well. In our experiments, we vary the number of mixture components, and select models according to classification accuracy. Observations are assigned to the class having the maximum *posterior* probability. Maximum A *Posteriori* (MAP) criterion aims at maximizing the *posterior* $p(\mathbf{C}|\mathbf{y})$ rather than the likelihood $p(\mathbf{y}|\mathbf{C})$.

4. EXPERIMENTAL DESIGN AND RESULTS

4.1. Experimental Design

The experiments were carried out on speech signals gathered from TIMIT database [16]. TIMIT collects reading speech from 630 native English speakers. There are totally 10 sentences from individual speaker, while each lasts approximately $3s$. Here we focused on phoneme classification. Each sentence has been manually labeled with phonetic symbols. There are 60 phonemes in total. In order to gather a sufficient amount of speech, we chose 46 speakers with equal gender partition, and speech signals covered all 60 phonemes, including vowels, fricatives, stops, affricates, nasals, semivowels and glides. To simplify the classification problem, we pre-grouped them into 3 large categories: vowels, fricatives and others. The unsupervised and supervised models were compared in a set of experiments: we stacked the basic time scale features into several longer time scales, and sparsified the stacked features with different degrees to test the consistency of the comparison. In the meanwhile of the performance comparison, we also anticipated to find out the role of time scales.

Following the preprocessing pipeline, we first extracted 25-d MFCCs from original speech signals with hamming windows in the time domain and triangular filters in the mel-frequency domain. Within these 25 dimensions, the so-called 0^{th} order MFCC was also included, which represents the log-energy of each short time frame. To investigate the role of time scales, we stacked the basic features into a variety of time scales, from $20ms$ scale up to $1100ms$ (20, 100, 150, 300, 500, 700, 900 and $1100ms$). Energy based sparsification was used afterwards. The degree of sparsification was controlled by thresholds leading to the retained energy from 100% to 65%. PCA was then carried out on stacked and sparsified features, and dimensionality of the features was reduced. For features having longer time scales than 20 ms, their dimensions were reduced to 100, and the dimension of the features at the basic time scale remained the same, i.e. 25.

After the preprocessing of features, we input the data into unsupervised and supervised models respectively. The training set covered 6 sentences from each of the 46 speakers, and the rest 4 sentences were used as test set. The ICA algorithm evaluated the unmixing matrix **W** of the training set, and the sources $\mathbf{S}^{train}$ were consequently recovered in unsupervised learning. Afterwards the sources were input to the naive Bayes classifier together with training set labels to estimate the parameters of the independent univariate Gaussians. For prediction, we preprocessed the test set following the same procedure. The **W** derived from the training set was applied to the test set to recover the sources $\mathbf{S}^{test}$. Whereafter naive Bayes classifier predicted the labels of the test set based on the test sources. We have used the exact same training and test set for the supervised model as for the unsupervised model, so as to exclude the comparison bias introduced by data. MoG models estimated a set of Gaussian distributions for each class from the training set, and fulfilled the label prediction on the test set. Both models provided us with a set of labels and a set of posterior label probabilities for both data sets.

4.2. Results

A set of experiments were carried out in 64 (8 times 8) different conditions, i.e. 8 time scales and 8 sparsification levels.

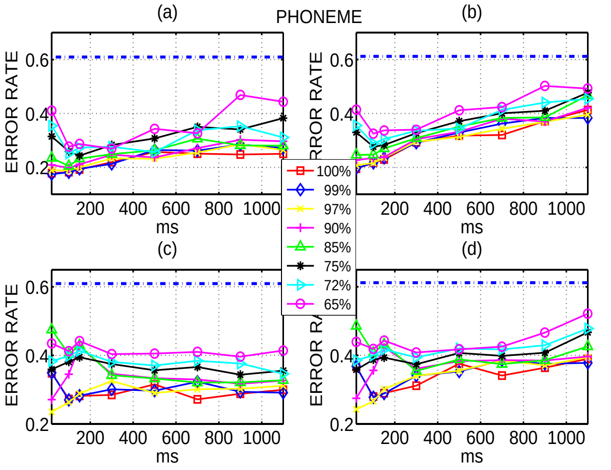

Fig. 3. Error rates as a function of time scales for different thresholds in phoneme classification. (a), (b): Training and test error rates of supervised MoG; (c), (d): Training and test error rates of unsupervised model, respectively; The 8 curves represent feature sparsification with retained energy from 100% to 65%. The dashed lines are the baseline error rates for random guessing. Results indicate that the relevant time scale is around the basic time scale.

Figure 3 presents the results of both supervised and unsupervised learning. The two plots (a) and (b) show the training and test error rates of the MoG models separately, whereas (c) and (d) are the training and test error rates of unsupervised learning (ICA+naive Bayes). The 8 curves in each panel represent the 8 EBS levels. First, it is quite obvious that features at longer time scales degraded the performance, which coincides with the conclusion from our previous research that phonemes are best modeled at short time scales [9, 17]. As we noticed, especially when retaining energy is 65%, high degree of sparsification decreased classification accuracy.

Error Rate Comparison From the above experiments we noticed that the performances of unsupervised and supervised models bear similarity w.r.t recognition error rates. To exam how well their representations are correlated, we measured the test performance of the resulting classifiers. High correlation between the error rates of the two schemes indicated similarity of the representations, shown in figure 4. The correlation is distinguished in phoneme classification task: for the given time scales and thresholds, data locate around $y = x$, and the correlation coefficient $\rho = 0.67, p < 1.38e - 009$.

Sample-to-Sample Correlation In order to reconfirm the finding and to account for the patterns of making decisions for both models, we followed the approach outlined above. We trained with the appropriate manual labels in supervised model to represent the human observer, and with the unsupervised -then- supervised learning scheme to represent the 'ecological' grouping. This experiment was also carried out on three groups of phonemes: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t, where eh stands for the vowel in the word 'BET', and ow for the vowel in 'BOAT'. Figure 5 presents the sample-to-sample classification results of both models. 25-d MFCCs were first sparsified, to keep 99% energy, and then PCA reduced the dimension to 6, and the resulting features were modeled by unsupervised and supervised learning methods separately. It is clear that two models had a similar pattern of making the correct prediction and making mistakes, and the percentage of matching (correct predictions from both models and misclassified samples from both models) between supervised and unsupervised learning was up to 91%.

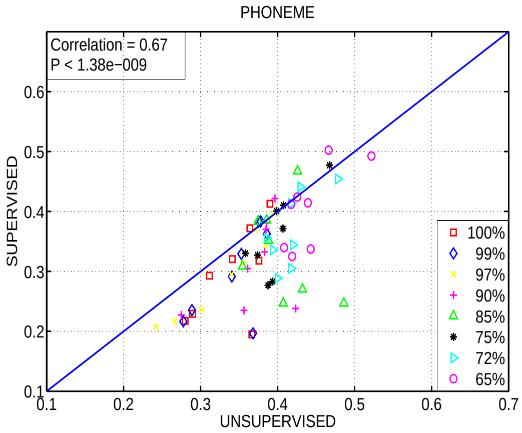

Fig. 4. Correlation between test error rates of supervised and unsupervised learning models. Solid lines indicate $y = x$. The correlation coefficient is 0.67.

Posterior Probability Comparison So far we have seen that there is a close correspondence at the level of error rates and sample-to-sample classification. A more detailed comparison can be obtained by considering the posterior probabilities obtained on a sample basis. We chose one experiment of the phoneme classification ($100ms$ time scale with 97% remaining energy) among the 64 experiments mentioned above. Figure 6 presents the posterior probability comparison of fricatives models. If two models are the exact match, we should expect that the posterior probabilities locate along the diagonal of the histograms with high distribution at $(1, 1)$ and $(0, 0)$. The matching in this case is around 57%.

5. CONCLUSION

With the purpose of understanding the exploitation of statistical regularities in human cognitive activity, we investigated the Cognitive Component Analysis. We have devised a protocol for testing the cognitive component hypothesis, that is to compare the performance of unsupervised learning, which aims at discovering statistical regularities, and supervised learning, which loosely represents human cognitive activity.

We have studies the COCA on phoneme level signals, and

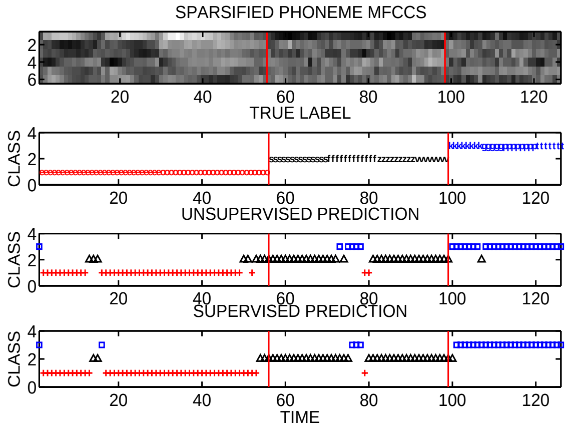

Fig. 5. Sample-to-sample phoneme classification among vowels, fricatives and stops. This first panel shows the temporal development of sparsified MFCCs. The boundaries of 3 phoneme classes are highlighted by vertical lines. Second panel gives the true labels, denoted by phonetic symbols. The last two panels give the unsupervised and supervised label predictions, marked by 3 shapes. The decision patterns of supervised and unsupervised learning show high similarity.

compared the performance of unsupervised and supervised learning at three levels: error rate level; sample-to-sample level; and the more detailed posterior probability level. In all the comparisons we have found evidence that supervised and unsupervised learning in fact do lead to similar representations.

6. ACKNOWLEDGEMENT

This work is supported by the Danish Technical Research Council, through the framework project ‘Intelligent Sound’ (STVF No. 26-04-0092), www.intelligentsound.org.

7. REFERENCES

[1] H.B. Barlow, “Unsupervised learning,” *Neural Computation*, vol. 1, pp. 295–311, 1989.

[2] A. J. Bell and T. J. Sejnowski, “The ‘independent components’ of natural scenes are edge filters,” *Vision Research*, vol. 37, pp. 3327–3338, 1997.

[3] P. Hoyer and A. Hyvrinen, “Independent component analysis applied to feature extraction from colour and stereo images,” *Network: Comput. Neural Syst.*, vol. 11, pp. 191–210, 2000.

[4] M. S. Lewicki, “Efficient coding of natural sounds,” *Nature Neuroscience*, vol. 5, pp. 356–363, 2002.

[5] L. K. Hansen, P. Ahrendt, and J. Larsen, “Towards cognitive component analysis,” in *AKRR’05 -International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

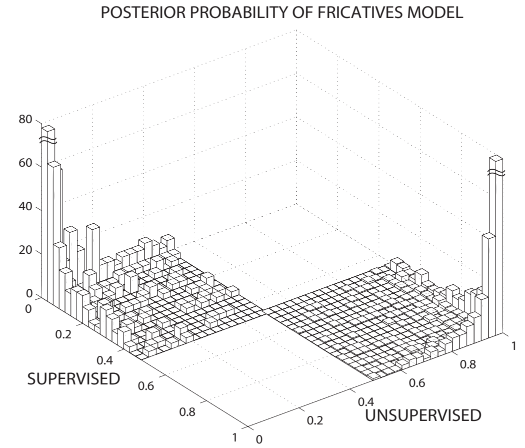

Fig. 6. Posterior probability comparison. Figure shows the histograms of the posterior probabilities provided by unsupervised and supervised fricatives models on the test set in the matching case. The two highest distributions locate at $(1, 1)$ and $(0, 0)$, which are 840 and 501 respectively.

[6] L. Feng and L. K. Hansen, “On low level cognitive components of speech,” in *Proc. International Conference on Computational Intelligence for Modelling*, 2005, vol. 2, pp. 852–857.

[7] D. J. Field, “What is the goal of sensory coding?,” *Neural Computation*, vol. 6, pp. 559–601, 1994.

[8] L. K. Hansen and L. Feng, “Cogito componentiter ergo sum,” in *Proc. ICA*, 2006, pp. 446–453.

[9] L. Feng and L. K. Hansen, “Phonemes as short time cognitive components,” in *Proc. ICASSP*, 2006, vol. 5, pp. 869–872.

[10] G. Mather, *Foundations of Perception*, Psychology Press, Hove, UK, 2006.

[11] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press Marketing, 2000.

[12] Y. Wang, Z. Liu, and J.C. Huang, “Multimedia content analysis using both audio and visual clues,” *IEEE Signal Processing Magazine*, vol. 17, 2000.

[13] D. Reisberg, *Cognition: Exploring the Science of the Mind*, W.W.Norton & Company, New York, USA, 2006.

[14] W. Kintsch, “Predication,” *Cognitive Science*, vol. 25, pp. 173–202, 2001.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*, OXFORD University Press, Oxford, UK, 1995.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, “The DARPA TIMIT acoustic phonetic continuous speech corpus CDROM,” in *NIST order number PB91-100354*, 1993.

[17] L. Feng and L. K. Hansen, “Cognitive components of speech at different time scales ,” in *Proc. CogSci*, 2007, pp. 983–988.