PHONEMES AS SHORT TIME COGNITIVE COMPONENTS

Ling Feng and Lars Kai Hansen
Informatics and Mathematical Modeling,
Technical University of Denmark, Denmark
lf@imm.dtu.dk, lkh@imm.dtu.dk

ABSTRACT

Cognitive component analysis (COCA) is defined as the process of unsupervised grouping of data such that the resulting group structure is well-aligned with that resulting from human cognitive activity [1]. In this paper we address COCA in the context short time sound features, finding phonemes which are the smallest contrastive units in the sound system of a language. Generalizable components were found deriving from phonemes based on homomorphic filtering features with basic time scale (20 msec). We sparsified the features based on energy as a preprocessing means to eliminate the intrinsic noise. Independent component analysis was compared with latent semantic indexing, and was demonstrated to be a more appropriate model in COCA.

1. INTRODUCTION

Cognitive component analysis (COCA) as a newly defined concept was first brought to bear in [1]: the process of unsupervised grouping of data such that the resulting group structure is well-aligned with that resulting from human cognitive activity. The concept is related to Lee and Seung's work on non-negative matrix factorization (NMF). In [2] they showed that components could be understood using concepts from gestalt theory: the factorization of an observation matrix in terms of a relatively small set of cognitive components leads to a parts-based object representation. In 2002, similar parts-based decompositions were obtained in a latent variable model based on non-negative linear mixtures of non-negative independent source signals [3]. Holistic, but parts-based, recognition of objects is frequently reported in perception studies across multiple modalities and increasingly in abstract data, where object recognition is a cognitive process.

The human perceptual system can model complex multi-agent scenery by using a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing agents. The fact motivating our interest in COCA is that representations found in human and animal perceptual systems closely resemble the theoretically optimal representations from the unsupervised signal separation, namely independent component analysis (ICA) [4, 5, 6]. This paper further discusses the generality of COCA based on the previous work [1, 7], and tries to answer the question: *Are such optimal representations based on abstract "independence" also relevant in higher cognitive functions?*

The phoneme is the smallest contrastive unit in the sound system of a language. Phoneme recognition is an active research field in speech recognition, see e.g., [8]. In [7] phonemes have been investigated by one of the generic tools of COCA analysis, namely Latent Semantic Indexing (LSI), and generalizable components and structures representing some of these smallest units have been found, as illustrated in Fig. 1. However whether the generalizable structure found in this work can assist phoneme recognition in general, still needs to be explored. Grouping by ICA has been pursued earlier for several abstract data types including text, dynamic text (chat), images, and combinations [9, 10, 11, 12, 13]. It was found that ICA is a more appropriate model than both LSI, which is too constrained, and clustering, which may in some instances be too flexible as a representation of text data.

The generality of ICA makes it possible to be utilized in many different areas. The classical application in signal processing of ICA model is blind source separation (BSS). A classical example of BSS is the cocktail party problem (CPP), see e.g., [14]. The problem is to separate the voices of different speakers, using recordings of one or more microphones. Comparing to BSS/CPP which is basically using original sound signals, the ICA model in COCA analysis applies on homomorphic filtering features, namely Mel-frequency Cepstral Coefficient (MFCC). MFCCs are short-term spectral features, and the mel-frequency warping transformation based on human auditive system. In COCA we are interested in a cognitive level, so to speak before semantics. The features we look for can be compared to the features a foreign speaker hears on entry. Sounds are recognized but without semantic reference. Hence, the cognitive context in our COCA is in the intermediate-level between source separation (low-level) and content recognition (high-level).

2. COGNITIVE COMPONENT ANALYSIS

2.1 Latent semantic indexing (LSI)

Latent semantic indexing is the PCA applied on abstract data such as text [15]. It is basically a tool for dimensionality reduction and also can be used to find group structure in data when the signal-to-noise ratio is high [7]. Our approach is inspired by LSI and the main innovation here is the active search for generalizable non-orthogonal linear features that may be described in terms of an independent component generative model.

A strong assumption in LSI is that the data have Gaussian distribution. Unfortunately, many real world data are *nongaussian*, instead very sparse [1, 7]. Hence LSI is often used as a tool to reduce dimensionality, which is post-processed to reveal cognitive components, e.g., by interactive visualization schemes [16].

2.2 Independent component analysis (ICA)

ICA algorithms can estimate independent components from linear mixtures [17], and has applications in many real world data. Here we discuss some basic characteristics of mixtures and the possible recovery of sources.

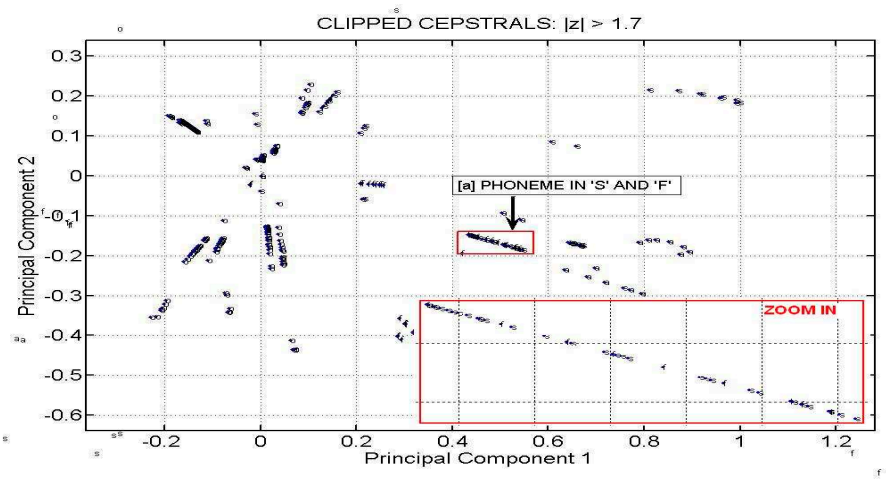

Fig. 1. Scatter plot of data on latent space

The latent space is formed by the two first principal components of the training data consisting of four separate utterances representing the sounds 's', 'o', 'f', 'a'. The structure clearly shows the sparse component mixture, with 'rays' emanating from the origin (0,0). The ray marked with an arrow contains a mixture of 's' and 'f' analysis windows, a generalizable characteristic feature associated with the vowel a-like sound that opens both an 's' and an 'f'.

First, we note that LSI/PCA is not able to reconstruct the mixing. PCA, being based on co-variance is simply not informed enough to solve the problem. To see this let the mixture be given as

$$\mathbf{X} = \mathbf{AS}, \qquad X_{j,t} = \sum_{k=1}^{K} A_{j,k} S_{k,t}, \tag{1}$$

where $X_{j,t}$ is the value of j'th feature in the t'th measurement, $A_{j,k}$ is the mixture coefficient linking feature j with the component k, while $S_{k,t}$ is the level of activity in the k'th source. In a text instance a feature is a term and the measurements are documents, while the components can be interpreted as topical contexts.

As a linear mixture is invariant to an invertible linear transformation we need to define a normalization of one of the matrices **A**, **S**. We do this by assuming that the sources are unit variance. As they are assumed independent the covariance will thus be trivially given as the unit matrix. LSI, hence PCA, of the measurement matrix is based on analysis of the covariance

$$\Sigma_X = \lim_{T\to\infty} \frac{1}{T} \mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}} \tag{2}$$

Clearly the information in $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is not enough to uniquely identify **A**, since if one solution **A** is found, any (row) rotated matrix $\widetilde{\mathbf{A}} = \mathbf{AU}$, $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ is also a solution, because $\widetilde{\mathbf{A}}$ has the same outer product as **A**. This is a potential problem for LSI based analysis. The ICA community has on the other hand devised many algorithms that use more informed statistics to locate **A** and thus **S**, see [17] for a recent review.

3. COMPONENT ANALYSIS FOR PHONEMES

The phoneme is defined as the class of sounds that are consistently perceived as representing a certain minimal linguistic unit in [18]. However phonologists have differing views of the phoneme, and two major ones are: in the American structuralist tradition, a phoneme is defined according to its allophones and environments; in the generative tradition, a phoneme is defined as a set of distinctive features [19]. An allophone is a phonetic variant of a phoneme in a particular language. According to the first view, the same phoneme can sound slightly different in different languages and environments. In American English approximately 40 phonemes are in use, of which 12 are vowels. Vowels vary in temporal duration between 40-400msec [18].

Four simple utterances 's', 'o', 'f', 'a' from the TIMIT database [20] were used for this demonstration. The basic time scale of 40 msec was used (windowing with 95% overlap), since the speech production system is generally considered stationary for time intervals on the order of 20-40 msec [18]. The windows were represented by 16 MFCCs. The temporal development of the mel-cepstral representation of the four utterances is presented in the upper panel of Fig. 4. After variance normalization we sparsified the energy based coefficients by zeroing windows of normalized magnitudes with a statistical $z < 1.4$, which retains 55% energy from original features. LSI/PCA was performed on the sparsified feature coefficients to get the most variant PCA components. The results from Fig. 1 seem to indicate that generalizable cognitive components corresponding to phonemes, e.g. /æ/ from utterance 's' and 'f', can be identified using linear component analysis. However the ray structures representing the phonemes are not aligned with the directions of the principal components, hence, an ICA scheme is required.

Six components ICA was applied on the PCA coefficients. Fig. 2 shows the scatter plot of sparsified features on the first two principal components derived from the 16 x 16 sparsified feature covariance matrix. The six independent sources were annotated as red circle, blue square, green diamond, magenta +, cyan triangle and black X respectively. The tags for the samples were labeled according to the independent sources, **S** matrix**,** from ICA analysis on sparsified and dimensionality reduced features. The arrows in Fig. 2 represent the directions of sources which are the column vectors of the mixing matrix **A** in equation (1). The 'ray' structure with rays emanating from the origin of the coordinate system is evident, and each ray along the vector belongs to one independent source. In order to testify the generalizability of this structure, a test set with another set of utterances 's', 'o', 'f', 'a' from TIMIT

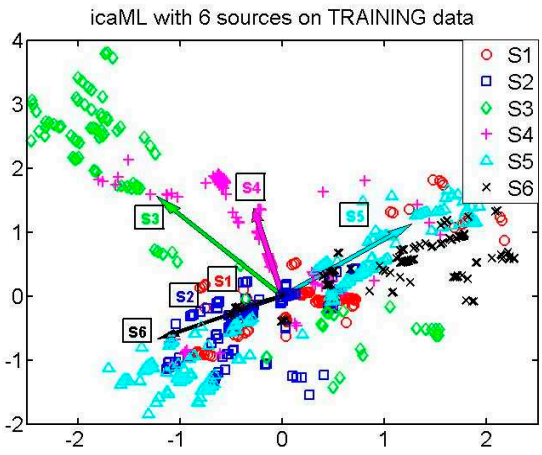

Fig. 2. Scatter plot of training data

Six components ICA performed on PCA coefficients. Scatter plot shows the data projected on the first two principal components derived from the sparsified features. The circle, square, diamond, +, triangle and X stand for 6 independent sources. The tags for the samples were labeled according to **S** matrix from ICA, and the arrows represent the directions of sources from mixing matrix **A**. The 'ray' structure with rays emanating from the origin (0,0) is evident.

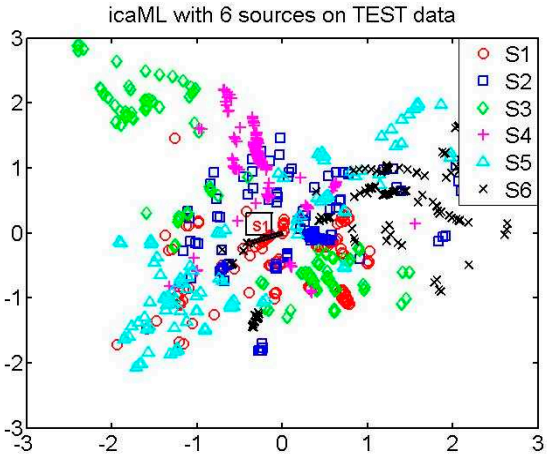

Fig. 3. Scatter plot of test data

Another set of utterances 's', 'o', 'f', 'a' was analyzed. The 'ray' structure is obvious and similar to the training set, emanating from the origin (0,0).

was analyzed using the same setup. The results are shown in Fig. 3. Here we only show the direction of the first source. Later we will demonstrate the cognitive content of this source.

Generalizability has been verified in another way by using two different implementations of ICA, namely maximum likelihood ICA (icaML) and the fast fixed-point algorithm for ICA (fastICA). IcaML algorithm is the estimation of the independent component as in the Infomax by Bell and Sejnowski [21] using a maximum likelihood formulation. Fig. 4 and 5 show the classification results from icaML and fastICA on training and test sets separately. In the two upper panels, the temporal development

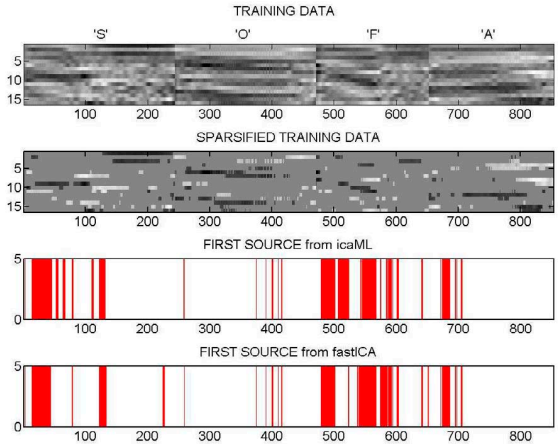

Fig. 4. MFCCs and Classification on Training set

In the two upper panels, the temporal development of the mel-frequency cepstral representations of the original 's', 'o', 'f', 'a' and 4 spasified ones is presented. The boundaries between them are clearly visible. 55% energy was retained after sparsification. The first independent sources from two ICA implementations are shown in the two lower panels: the vertical lines indicate the locations of windows belonging to the first source. Results from two ICA algorithms are similar. A large percentage of the windows locate in, approximately, windows No. 1 to No. 133 for 's', and No. 471 to No. 600 for 'f'. It indicates the feature is related to the similar /æ/ sound that opens both 's' and 'f'.

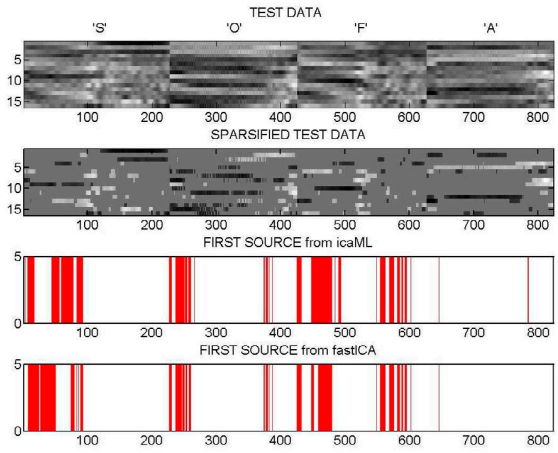

Fig. 5. MFCCs and Classification on Test set

The two upper panels show the temporal development of the mel-frequency cepstral representations of the four original utterances and four spasified ones. 60% energy was left for test set. The two lower panels show the first independent sources from icaML and fastICA: the vertical lines indicate the locations of windows belonging to the first source. Two panels look quite similar. The similar scenario shown in Fig. 4 for training set happened again on test set, which indicates the feature is related to the similar /æ/ sound that opens both 's' and 'f'. However there are more mis-detections located outside the above ranges.

of the mel-frequency cepstral representations of the four original utterances and four sparsified utterances is presented with the sequence of 's', 'o', 'f', 'a'. The boundaries between the four utterances are clearly visible, and the utterances show much similarity between the two samples (test and train), however, they are of quite different duration. For training set, 55% energy was retained after sparsification; and 60% energy was left for test set. The first independent sources from two ICA algorithms are shown in the two lower panels of Fig. 4 and 5: the vertical lines indicate the locations of windows belonging to the first source. It is quite clear that the results of icaML resemble those of fastICA. For training set, we notice that a large percentage of the windows locate in the first part of 's' and 'f' utterances, which approximately from windows No. 1 to No. 133 for 's', and No. 471 to No. 600 for 'f'. It indicates the feature is related to the similar /æ/ sound that opens both 's' and 'f'. A similar scenario happened in test set, however there are more lines locate outside the above ranges. Our interpretation is the windows containing low energy (almost zero) have simply been classified into the first class. The classification has been improved while we slightly reduced the threshold for sparsification. However low threshold brings more noise, which increases the classification error.

4. CONCLUSION

The generality of cognitive component analysis, which is 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 in this paper. We posit speech COCA in a cognitive level before semantics. In other words, sounds (sources) are recognizable, but without semantic reference. Therefore COCA is localized in the intermediate-level between source separation (low-level) and content recognition (high-level).

We have studied the derived cognitive components of phonemes from short time homomorphic filtering features with energy based sparsification. ICA on short-term spectral features, MFCC, was compared with latent semantic indexing, and was demonstrated to be a more appropriate model in COCA.

The fact that we find the 'ray' structure of cognitively relevant components by simple unsupervised learning based on sparse linear component analysis highlights the possibility of using unlabeled samples in supervised learning.

5. ACKNOWLEDGMENT

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

REFERENCES

[1] 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*. Jun 2005, Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society.

[2] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] Pedro A. D. F. R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen, "Mean-field approaches to independent component analysis," *Neural Comput.*, vol. 14, no. 4, pp. 889–918, 2002.

[4] Anthony J. Bell and Terrence J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[5] Patrik Hoyer and Aapo Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Comput. Neural Syst.*, vol. 11, no. 3, pp. 191–210, 2000.

[6] M.S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[7] L. Feng and L. K. Hansen, "On low level cognitive components of speech," *accepted* in CIMCA'05 -International Conference on Computational Intelligence for Modelling, Nov 2005.

[8] Ofer Dekel, Joseph Keshet, and Yoram Singer, "An online algorithm for hierarchical phoneme classification," in *MLMI*, pp. 146–158, 2004.

[9] L. K. Hansen, J. Larsen, and T. Kolenda, "On independent component analysis for multimedia signals," in *Multimedia Image and Video Processing*, pp. 175–199. CRC Press, Sep 2000.

[10] L. K. Hansen, J. Larsen, and T. Kolenda, "Blind detection of independent dynamic components," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, vol. 5, pp. 3197–3200, 2001.

[11] T. Kolenda, L. K. Hansen, and J. Larsen, "Signal detection using ICA: Application to chat room topic spotting," in *Third International Conference on Independent Component Analysis and Blind Source Separation*, pp. 540–545, 2001.

[12] T. Kolenda, L. K. Hansen, J. Larsen, and O.Winther, "Independent component analysis for understanding multimedia content," in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, H. Bourlard et al. Ed., Piscataway, New Jersey, 2002, pp. 757–766, IEEE Press, Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[13] J. Larsen, L.K. Hansen, T. Kolenda, and F.AA. Nielsen, "Independent component analysis in multimedia modeling," in *Fourth International Symposion on Independent Component Analysis and Blind Source Separation*, Shun ichi Amari et al. Ed., Nara, Japan, apr 2003, pp. 687–696, Invited Paper.

[14] S. Haykin and Z. Chen, "The Cocktail Party Problem," Neural Comp. 2005, vol. 17, pp. 1875-1902.

[15] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[16] T.K. Landauer, D. Laham, and M. Derr, "From paragraph to graph: latent semantic analysis for information visualization," *Proc Natl Acad Sci*, vol. 101, no. Sup. 1, pp. 5214–5219, 2004.

[17] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

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

[19] E. E. Loos, S. Anderson, D. H. Jr. Day, P. C. Jordan and J. D. Wingate, "Glossary of linguistic terms," SIL International, 2004. http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/index.htm

[20] J. S. Garofolo et al., *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, NIST, 1993.

[21] A. Bell and T.J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Neural Comp. 1995, vol.7, pp. 1129-1159.