# ON LOW-LEVEL COGNITIVE COMPONENTS OF SPEECH

*Ling Feng, Lars Kai Hansen, and Jan Larsen*

Intelligent Signal Processing,
Informatics and Mathematical Modelling,
Technical University of Denmark B321,
DK-2800 Kgs. Lyngby, Denmark

**ABSTRACT**

In this paper we analyze speech for low-level cognitive features using linear component analysis. We demonstrate generalizable component 'fingerprints' stemming from both phonemes and speaker. Phonemes are fingerprints found at the basic analysis window time scale (20 msec), while speaker "voiceprints" are found at time scales around 1000 msec. The analysis is based on homomorphic filtering features and energy based sparsification.

## 1. INTRODUCTION

The human perceptional system can model complex multi-agent scenery. It is well documented that humans use a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing agents, such as speakers, gestures, affections etc. Such unsupervised signal separation has also been achieved in computers using a variety of independent component analysis algorithms, see e.g., [1]. It is an intriguing fact that representations are found in human and animal perceptual systems which closely resembles the information theoretically optimal representations obtained by independent component analysis, see e.g., [2] on visual contrast detection, [3] on visual features involved in color and stereo processing, and [4] on representations of sound features.

In a companion paper presented at this meeting [5], we investigate the independent *cognitive* component hypothesis, which basically asks the question of humans use similar methods in basic perception as well as in more generic and abstract data. We denote in the companion paper algorithms that present spontaneous cognition as a result of unsupervised learning as *cognitive component analysis* (COCA).

Here we are interested in pursuing this idea in the context of speech. We are interested in the auditory aspects, not contents. We will focus on two aspects, phoneme features and speaker features. As in the companion paper, our presentation will be qualitative, mainly be based on simple visualizations of data, thus we avoid unnecessary algebraic complication.

Grouping of events or objects in more or less distinct categories is fundamental to human cognition. In machine learning, classification is a rather well-understood task when based on *labelled* examples [6]. In this case classification belongs to the class of *supervised* learning problems. Clustering is a closely related *unsupervised* learning problem, in which we use general statistical rules to group objects, without a priori providing a set of labelled examples. It is a fascinating finding in many real world data sets that the label structure discovered by unsupervised learning closely coincides with labels obtained by letting a human or a group of humans perform classification, labels derived from human cognition. We have earlier pursued grouping by independent component analysis in several abstract data types including text, dynamic text (chat), images, and combinations hereof, see e.g., [7, 8, 9, 10, 11]. It was found in this work that independent component analysis is a more appropriate model than both principal component analysis (which is too contrained) and clustering, which may in some instances be too flexible, say as a representation of text data.

## 2. COGNITIVE COMPONENT ANALYSIS

In 1999 Lee and Seung introduced the method of non-negative matrix factorization (NMF) [12] as a scheme for parts-based object recognition[1]. They argued that the factorization of an observation matrix in terms of a relatively small set of cognitive components, each consisting of a non-negative feature vector and a non-negative activation vector leads to a parts based object representation. They demonstrated the values of the non-negative representation for objects in images and in text. More recently, in 2002, it was shown that very similar parts-based decompositions were obtained in a latent variable model based on positive linear mixtures of positive *independent* source signals [13]. 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. Together these findings are often referred to as instances of the more general *Gestalt laws*.

### 2.1. Latent semantic indexing (LSI)

Principal component analysis (PCA) has been (re-)invented over and over in virtually all branches of sciences. Analysis of variance is a very useful tool for dimensional reduction of high dimensional correlated data and may be

[1] This section is based in part on the companion paper [5]

used to find group structure in data when the signal-to-noise ratio is high. PCA has been used for basic perceptual feature analysis, such as in images under the name Karhunen-Loeve transform see, e.g., [14], and for analysis of abstract data such as text under the name latent semantic indexing (LSI) [15]. Our approach is inspired by LSI and the main innovation in our discussion is the active search for generalizable non-orthogonal linear features that may be described in terms of an independent component generative model.

Salton proposed the so-called vector space representation for statistical modeling of text data, for a review see [16]. A term set is chosen and a document is represented by the vector of term frequencies. A document database then forms a so-called term-document matrix. The vector space representation can be used for classification and retrieval by noting that similar documents are somehow expected to be 'close' in the vector space. A metric can be based on the simple Euclidean distance if document vectors are properly normalized, otherwise angular distance may be useful. This approach is principled, fast, and language independent. Deerwester and co-workers developed the concept of latent semantics based on principal component analysis of the term-document matrix [15]. The fundamental observation behind the latent semantic indexing (LSI) approach is that similar documents are using similar vocabularies, hence, the vectors of a given topic could appear as produced by a stochastic process with highly correlated term-entries. By projecting the term-frequency vectors on a relatively low dimensional subspace, say determined by the maximal amount of variance one would be able to filter out the inevitable 'noise'. Noise should here be thought of as individual document differences in term usage within a specific context. For well-defined topics, one could simply hope that a given context would have a stable core term set that would come out as a 'direction' in the term vector space. Below we will explain why this is likely not to happen in general document databases, and LSI is therefore often used as a dimensional reduction tool, which is then post-processed to reveal cognitive components, e.g., by interactive visualization schemes [17].

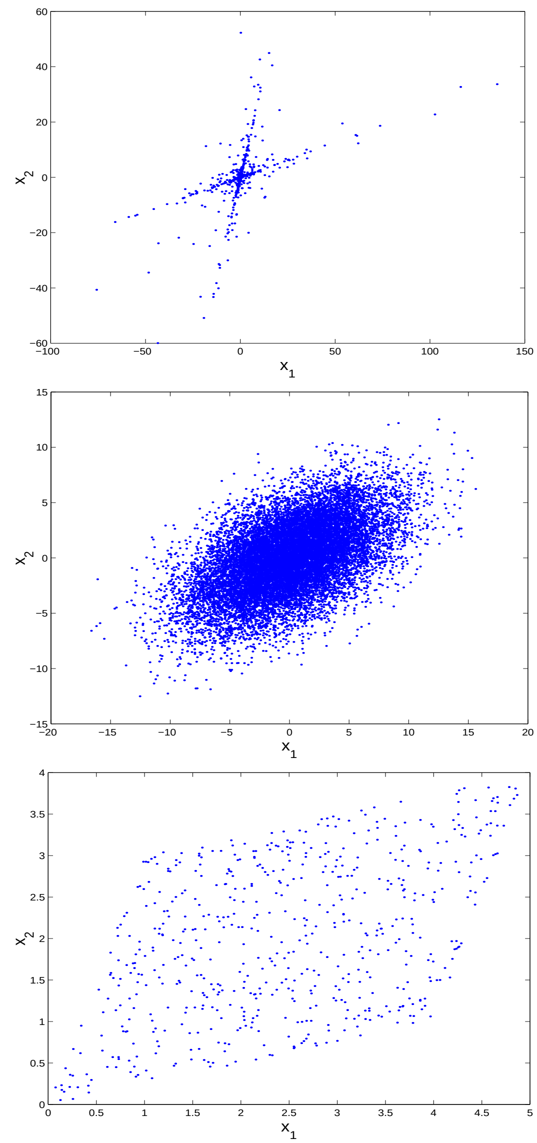


Figure 1. Prototypical feature distributions produced by a linear mixture, based on sparse (top), normal (middle), or dense source signals (bottom), respectively. The characteristic of the sparse signal is that it consists of relatively few large magnitude samples on a background of small signals.

### 2.2. Non-negative matrix factorization (NMF)

Noting that a non-negative decomposition could lead to a parts-based decomposition, Lee and Seung analyzed several data sets using the NMF decomposition technique [12]. Non-uniqueness of the components is a major challenge for this method and has been discussed in detail by Donoho and Stodden [18]. A possible route to more unique solutions, hence, potentially more interpretable and relevant components is to add a priori knowledge, e.g., in form of independence assumptions. An algorithm for reconstruction of positive independent components from a positive mixture is discussed in [13].

### 2.3. Independent component analysis (ICA)

Blind signal separation is the general problem of recovering source signals from an unknown mixture. This aim is in general not feasible without additional information. If we assume that the unknown mixture is linear, i.e., that the mixture is a linear combination of the sources, and furthermore assume that the sources are statistically independent processes it is often possible to recover sources and mixing, using a variety of independent component analysis techniques [1]. Here we will discuss some basic characteristics of mixtures and the possible recovery of sources.

First, we note that LSI/PCA is not able to reconstruct the mixing, PCA, being based on co-variance is simply

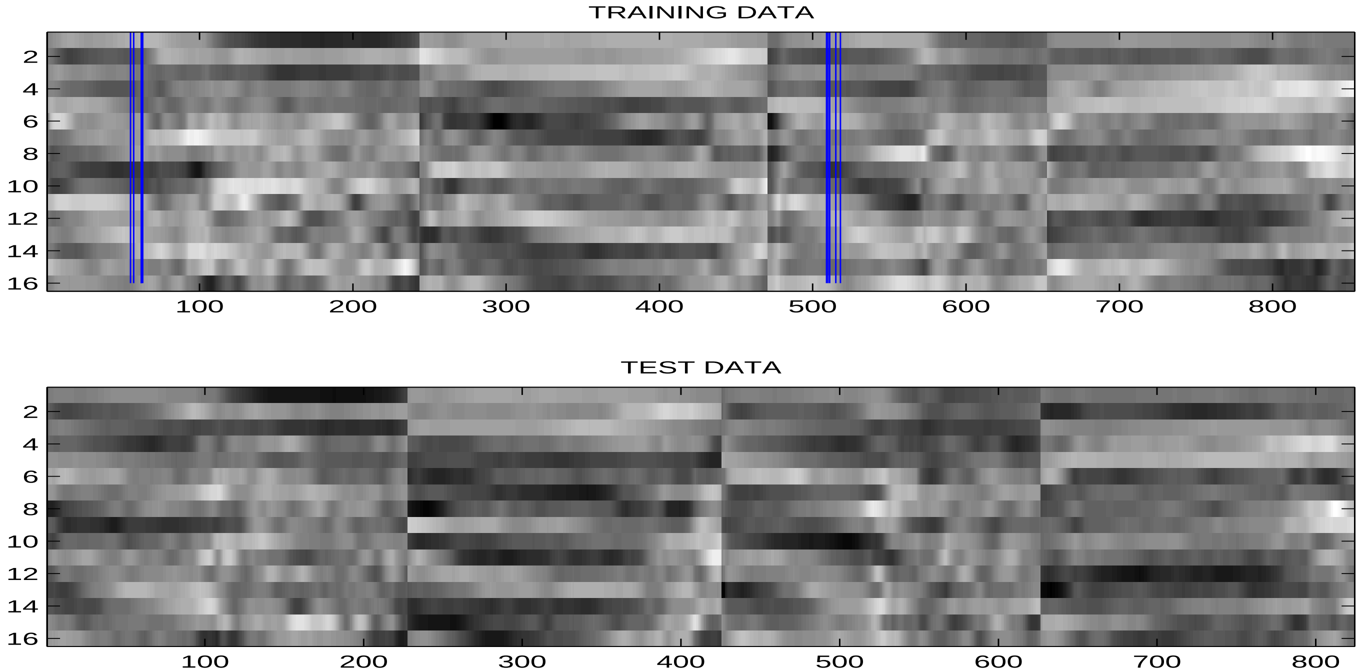


Figure 2. Four separate utterances are concatenated for this experiment, representing the sounds ‘s’, ‘o’, ‘f’, ‘a’. Each concatenated set of utterances is represented twice: In a training set and in a test set. The cepstral coefficient sequences for the two sets are shown in the two panels. The boundaries between the four utterances are clearly visible, and we note that the utterances show much similarity between the two samples (test and train), however, they are of quite different duration. The first of the two phones of the utterance ‘s’ is the opening a-like phoneme. In the upper panel we have added a set vertical lines to indicate positions of analysis windows that belong to a generalizable finger print feature further discussed in figure 3

not informed enough to solve the problem. To see this Let the mixture be given as

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \quad X_{j,t} = \sum_{k=1}^{K} A_{j,k} S_{k,t}, \tag{1}$$

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, the components are best thought as topical contexts. The $k$’th column $A_{j,k}$ holds the relative frequencies of term occurrence in documents within context $k$. The source matrix element $S_{k,t}$ quantifies the level of expression of context $k$ in document $t$.

As a linear mixture is invariant to an invertible linear transformation we need define a normalization of one of the matrices $\mathbf{A}, \mathbf{S}$. We will do this by assuming that the sources are unit variance. As they are assumed independent the covariance will be trivial,

$$\Sigma_S = \lim_{T\to\infty} \frac{1}{T} \mathbf{S}\mathbf{S}^\top = \mathbf{I}. \tag{2}$$

LSI, hence PCA, of the measurement matrix is based on analysis of the covariance

$$\Sigma_X = \lim_{T\to\infty} \frac{1}{T} \mathbf{X}\mathbf{X}^\top = \mathbf{A}\mathbf{A}^\top. \tag{3}$$

Clearly the information in $\mathbf{A}\mathbf{A}^\top$ is not enough to uniquely identify $\mathbf{A}$, since if a solution $\mathbf{A}$ is found, any (row) rotated matrix $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{U}, \mathbf{U}\mathbf{U}^\top = \mathbf{I}$ is also a solution, because $\tilde{\mathbf{A}}$ has the same outer product as $\mathbf{A}$.

This is a potential problem for LSI based analysis. If the document database can be modelled as in eq. (1) then the original characteristic context histograms will not be found by LSI. The field of independent component analysis has on the other hand devised many algorithms that use more informed statistics to locate $\mathbf{A}$ and thus $\mathbf{S}$, see [1] for a recent review.

The histogram of a source signal can roughly be described as sparse, normal, or dense. Scatter plots of projections of mixtures drawn from source distributions with one of these three characteristics are shown in Figure 1. In the upper panel of Figure 1 we show the typical appearance of a sparse source mixture. The sparse signal consists of relatively few large magnitude samples in a background of a large number of small signals. When mixing such independent sparse signals as in Eq. (1), we obtain a set of rays emanating from origo. The directions of the rays are directly given by the column vectors of the $\mathbf{A}$-matrix.

If the sources are truly normal distributed like in the middle panel of Figure 1, there is no additional information but the covariance matrix. Hence, in some sense this is a singular worst case for separation. Because we work from finite samples an ICA method, which assumes some

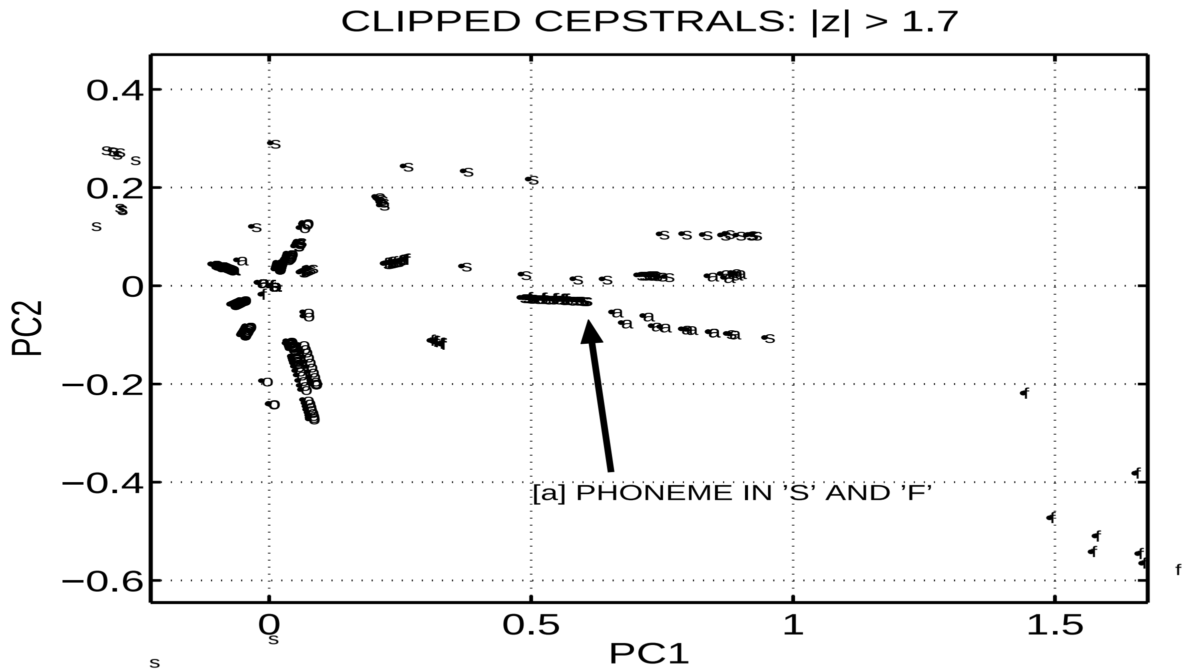


Figure 3. We show the latent space formed by the two first principal components of the training data consisting of four separate utterances shown in figure 2 representing the sounds 's', 'o', 'f', 'a'. The structure clearly resembles the sparse component mixture in figure 1, with 'rays' emanating from the origin $(0, 0)$. The ray marked with an arrow contains a mixture of 's' and 'f' analysis windows. The locations of these window were indicated by vertical lines in figure 2. This feature also contains a mixture of windows from both the training and test utterances, hence, is a generalizable characteristic feature associated with the vowel a-like sound that opens both an 's' and an 'f'.

non-normality, will in fact often find good approximations to the mixing matrix, simply because a finite normal sample will have non-normal oddities. But fortunately, many, many interesting real world data sets are not anywhere near normal, rather they are typically very sparse, hence, more similar to the upper panel of Figure 1.

## 3. COMPONENT ANALYSIS OF SPEECH

In the authoritative textbook 'Discrete-Time Processing of Speech Signals' by Deller et al. [19] the phoneme is defined as the class of sounds that are consistently perceived as representing a certain minimal linguistic unit. In American English approximately 40 phonemes are in use, of which 12 are vowels. Vowels vary in temporal duration between 40-400msec [19].

The processes in the speech production system are generally considered stationary for time intervals on the order of 20 msec [19], hence, we will use an analysis window of this duration. In each window we represent the sound signal, i.e., 200 signal values for a sampling rate of 10kHz, by a relatively low-dimensional feature vector. This feature vector is obtained by homomorphic filtering, as often invoked in speech recognition. The resulting, so-called *cepstral coefficients* are designed to reduce the influence of the speech pitch, i.e., the speaker's 'tone' [19]. The cepstral coefficients are used in speaker independent speech recognition, because in this context the pitch is a confound. The cepstral coefficients are supposed to emphasize the linguistic content and suppress the speakers 'voice print'.

A small set of four simple utterances ('s', 'o', 'f', 'a') from the TIMIT database [21] was used for this demonstration. For the analysis we used 20 msec analysis windows with 50% overlap. The windows were represented by 16 cepstral coeffients. The temporal development of the cepstral representation of the four utterances is presented in two versions in figure 2, in the upper panel for the training set, and in the lower panel for a test set. After variance normalization we sparsified the coefficients by zeroing windows of normalized magnitudes less than $z > 1.7$. In figure 3 we show the scatter plot of the set of windows projected onto the first two principal components derived from the $16 \times 16$ feature covariance matrix. There is a marked 'ray' structure with rays emanating from the origin of the coordinate system $(0, 0)$. The projected features from the set of analysis windows have been annotated with their utterance origin. The arrow points to a linear ray structure which contains windows from utterances 's' and 'f'. In order to understand which part of the utterances these windows belongs to, we have marked up several points (windows) in 3 and we have indicated the temporal location of these windows as vertical stripes in figure 2 . From this it is clear that the feature is related to the similar a-like sound that opens both 's' and 'f'.. The generalizability of this structure was proven by creating a similar plot with the projections of the test set windows

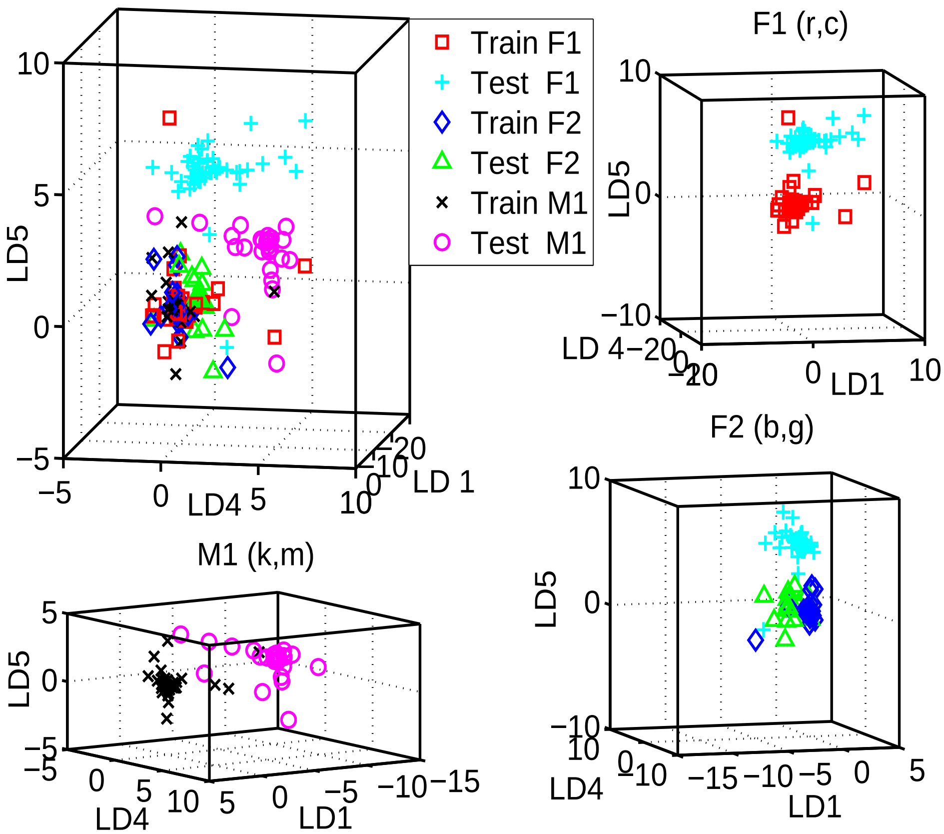


Figure 4. Text *dependent* speaker recognition. This analysis is based on a subset of the speakers enrolled in the ELSDSR speaker recognition database [20]. Here we focus on text dependent speech, hence, the same set of sentences is used for all speakers. The basic analysis window of the speech signal is represented by $12$ mel-cepstral coefficients (MFCC's) for this experiment. Fifty basic analysis windows are concatenated to form an intermediate time scale representation to capture human voice identity. The dimensionality of the aggregate representation is thus $50 \times 12$. After variance normalization we sparsified the coefficients by retaining the upper 1% magnitude fraction. We used a training set based on speech from three speakers (annotated as: F1 (female, red square) F2 (female, blue diamond) and M1 (male, black x); and corresponding test sets: F1 (cyan +), F2 (green triangle) and M1 (magenta circle). The data from the training set is submitted for principal component analysis, we show the scatter plots of both training data and test data in the space of a few principal or latent components. In the upper left display all data points are shown as represented in the space of the first, fourth, and fifth principal (latent) components. There is an evident ray structure corresponding to a generative ICA model based on linear mixing of sparse sources, i.e., similar to the situation seen at the time scale of the basic analysis window (20 msec). It appears that the structure is indeed speaker dependent in the sense that the ray systems are offset from the origin. We conclude that for this text dependent representation we find a mixture of phoneme like features and speaker identity features.

(data not not shown). This structure is indeed generalizable in constrast to some of the other ray-like structures that apparently are too instance specific to provide generalization from the relative small set of training data.

The results seem to indicate that generalizable cognitive components corresponding to phonemes can be identified using linear component analysis. The ray structures representing the phoneme is not aligned with the directions of the principal components, hence, an independent component analysis scheme is required. Phoneme recognition is an active research field in speech recognition, see e.g., [22], and it is an interesting issue for further research whether the generalizable structure found in this work can assist phoneme recognition in general.

## 4. VOICE PRINT COMPONENTS

While phonemes are universal components of language and generalizable in large populations, *speaker identity* plays an important in both social contexts and in speech based engineering applications, for example related to access control, see e.g., [23].

Speaker recognition has two aspects: Speaker identi-

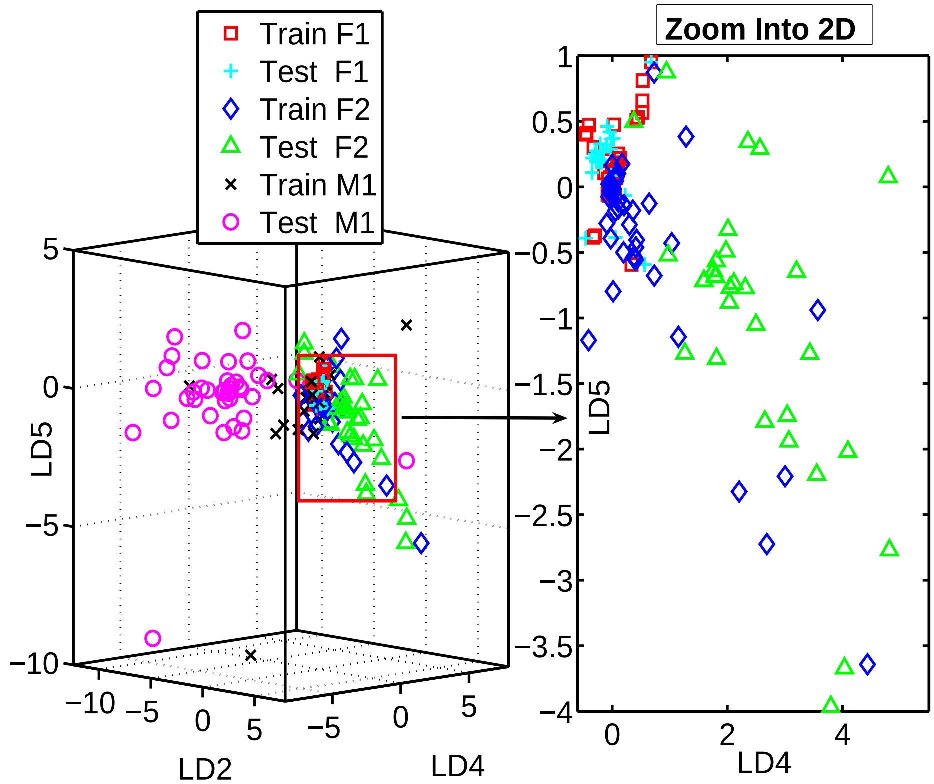


Figure 5. Text *independent* speaker recognition. This analysis is based on a subset of the speakers enrolled in the ELSDSR speaker recognition database [20]. Here we focus on text independent speech, hence, each subject is enrolled with a different set of sentences. The basic analysis window of the speech signal is represented by $12$ mel-cepstral coefficients (MFCC's) for this experiment. Fifty basic analysis windows are concatenated to form an intermediate time scale representation to capture human voice identity. The dimensionality of the aggregate representation is thus $50 \times 12$. After variance normalization we sparsified the coefficients by retaining the upper 1% magnitude fraction. We used a training set based on speech from three speakers (annotated as: F1 (female, red square) F2 (female, blue diamond) and M1 (male, black x); and corresponding test sets: F1 (cyan +), F2 (green triangle) and M1 (magenta circle). As in the previous experiment the data from the training set is submitted for principal component analysis, and we show the scatter plots of both training data and test data in the space of a few principal or latent components. In the left panel all data points are shown as represented in the space of the second, fourth, and fifth latent components. There is an evident ray structure corresponding to a generative ICA model based on linear mixing of sparse sources. In contrast to the text independent case above we see that the ray structure is solely determined by the speaker identity. The right hand side plot shows a close up of the structure for the female speaker F2: emphasizing the generalizability. The rays from the training and test sets are closely aligned.

fication, and speaker verification. Speaker verification is the process of determining whether a postulated speaker identity is correct, while speaker identification is the process of finding the identity of an unknown speaker by comparing his/her voice with all the registered/known speakers in the database [24]. Compared to verification, identification is more complicated. In the case that the unknown speaker must come from a fixed set of enrolled speakers, the system is referred to as a closed-set system. Moreover speaker recognition systems are divided according to the spoken text modality: text-dependent and text-independent. Compared to text-dependent speaker recognition, text-independent systems are more flexible, but also more complex. Feature extraction is very important for speaker recognition systems. The most widely accepted features for speaker recognition are mel-frequency cepstral coefficients (MFCC). The MFCC's are perceptually weighted cepstral coefficients [19].

According to our basic hypothesis the speaker dependent generalizable 'cognitive' components should be elucidated by Latent Semantic Analysis (LSA). To test the hypothesis we study here three speakers' voice messages from our in-house ELSDSR speech database [20]. In this database, read text is recorded using a MARANTZ PMD670 portable solid state recorder, and stored in PCM (wav) format. The sampling frequency was 16 kHz. ELSDSR contains voice messages from a total of 22 speakers (12M/ 10 F) of age from 24y to 63y.

Speaker identity information in speech can be categorized into a hierarchy ranging from low-level cues, such as the basic sound of a person's voice, which is related to physical traits of the vocal apparatus, to high-level cues, such as particular word usage (idiolect), conversational patterns and even topics of conversations, which is related to learned habits and style [25].

For the first *text dependent* speaker recognition experiment, signals from speakers F1, F2 and M1 reading the same text content were selected, and divided into training set (52.5sec) and test set (35.5sec). The frames with 20msec signal content were blocked without overlap, and 12 MFCC's were extracted from each frame. To form the long-term features, 50 basic analysis windows were concatenated. The total number of such expanded frames in the analysis was 522. Energy based sparsification was performed on the high dimensional data, and the upper 1% fraction was retained. Finally, LSA (PCA) was performed on the sparsified data to get the scatter plot of the data on the subspace spanned by three latent dimensions (LD), shown in figure 4. We annotated the data points for the training set of the three speakers as: F1 (red square) F2 (blue diamond) and M1 (black x); and test set as: F1 (cyan +), F2 (green triangle) and M1 (magenta circle).

Since the speakers read the same text content (training and test set are different) the red, green and black points emanate from (0,0), and show similar sparse ICA 'ray' structures. Since we are here using the same texts for all speakers, these features can be thought of a characteristic of the given words, i.e., similar to the phoneme features found above. However, importantly the rays also show speaker-dependent characteristics. This is most easily appreciated by inspecting the three plots to the right. Here the situations for the individual speaker are depicted, as seen the features do not generalize in a simple way, it appears that there is an offset between test and training data, which is speaker dependent, we stipulate that this effect is an interaction between the text content and the speaker identity.

We now turn to *text independent* speech. We study the same three speakers as before, two female and one male. The representation is identical to the one used for the text dependent experiment. The scatter plot of test and training data is shown in 3D subspace based on latent dimensions 2, 4 and 5. The data points of the three speakers are annotated as as text dependent case. Figure 5 shows that data points from 2 female speakers and the male speaker are aligned for both training and test set. The right side panel shows a zoomed in and projected subset of the data belonging to the two female speakers in latent dimension 4 and 5. The 'ray' structure emanates from (0,0) without offsets.

## 5. CONCLUSION

We have proposed to define cognitive component analysis as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. In this paper we have studied the so derived cognitive components of speech signals. We used homomorphic filtering to derive features and we analyzed the excursion set after thresholding based on energy.

At short time scales, we found generalizable features corresponding to phonemes. Phonemes are universal linguistic atoms recognized by large populations.

Humans swiftly and reliably recognize other another human's voice. We have shown that at intermediate time scales, 500-1000msec, there are generalizable speaker specific sparse components.

The fact that we find such cognitively relevant component by simple unsupervised learning based on sparse linear component analysis lends further support to our working hypothesis that humans could use such information theoretical representations, not only in basic perception tasks, but also when analyzing more abstract data.

## 6. ACKNOWLEDGMENTS

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

## 7. REFERENCES


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

[2] 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.

[3] 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.

[4] M.S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, 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*. jun 2005, Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society.

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

[7] 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.

[8] 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*, 2001, vol. 5, pp. 3197–3200.

[9] 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*, 2001, pp. 540–545.

[10] 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.

[11] 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.

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

[13] 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.

[14] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[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] Gerard Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.*, Addison-Wesley, 1989.

[17] 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.

[18] David L. Donoho and Victoria Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *NIPS*, 2003.

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

[20] *http://www.imm.dtu.dk/∼lf/ELSDSR.htm*, 2005.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, NIST, 1993.

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

[23] *http://www.research.ibm.com/VIVA_Demo*, 2005.

[24] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *ICASSP 2002*, 2002.

[25] J.P. Campbell, D.A. Reynolds, and R.B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proceedings of Eurospeech-2003 (Geneva, Switzerland)*, 2003, pp. 2665–2668.