Cognitive Components of Speech at Different Time Scales

Ling Feng (lf@imm.dtu.dk)
Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs. Lyngby, Denmark

Lars Kai Hansen (lkh@imm.dtu.dk)
Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs. Lyngby, Denmark

Abstract

Cognitive component analysis (COCA) is defined as unsupervised grouping of data leading to a group structure well-aligned with that resulting from human cognitive activity. We focus here on speech at different time scales looking for possible hidden 'cognitive structure'. Statistical regularities have earlier been revealed at multiple time scales corresponding to: phoneme, gender, height and speaker identity. We here show that the same simple unsupervised learning algorithm can detect these cues. Our basic features are 25-dimensional short-time Mel-frequency weighted cepstral coefficients, assumed to model the basic representation of the human auditory system. The basic features are aggregated in time to obtain features at longer time scales. Simple energy based filtering is used to achieve a sparse representation. Our hypothesis is now basically ecological: We hypothesize that features that are essentially independent in a reasonable ensemble can be efficiently coded using a sparse independent component representation. The representations are indeed shown to be very similar between supervised learning (invoking cognitive activity) and unsupervised learning (statistical regularities), hence lending additional support to our cognitive component hypothesis.

Keywords: Cognitive component analysis; time scales; energy based sparsification; statistical regularity; unsupervised learning; supervised learning.

Introduction

The evolution of human cognition is an on-going interplay between statistical properties of the ecology, the process of natural selection, and learning. Robust statistical regularities will be exploited by an evolutionary optimized brain (Barlow, 1989). Statistical independence may be one such regularity, which would allow the system to take advantage of factorial codes of much lower complexity than those pertinent to the full joint distribution. In (Wagensberg, 2000), the success of given 'life forms' is linked to their ability to recognize independence between predictable and un-predictable process in a given niche. This represents a precision of the classical Darwinian paradigm by arguing that natural selection simply favors innovations which increase the independence of the agent and un-predictable processes. The agent can be an individual or a group. The resulting human cognitive system can model complex multi-agent scenery, and use a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing processes.

The optimized representations for low level perception are indeed based on independence in relevant natural ensemble statistics. This has been demonstrated by a variety of independent component analysis (ICA) algorithms, whose representations closely resemble those found in natural perceptual systems. Examples are, e.g., visual features (Bell & Sejnowski, 1997; Hoyer & Hyvrinen, 2000), and sound features (Lewicki, 2002).

Within an attempt to generalize these findings to higher cognitive functions we proposed and tested the independent cognitive component hypothesis, which basically asks the question: *Do humans also use information theoretically optimal ICA methods in more generic and abstract data analysis?* Cognitive component analysis (COCA) is thus simply defined as the process of unsupervised grouping of abstract data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity (Hansen, Ahrendt, & Larsen, 2005). For the preliminary research on COCA, human cognitive activity is restricted to the human labels in supervised learning methods. This interpretation is not comprehensive, however it is capable of representing some intrinsic mechanism of human cognition. Further more, COCA is not limited to one specific technique, but rather a conglomerate of different techniques. We envision that efficient representations of high level processes are based on sparse distributed codes and approximate independence, similar to what has been found for more basic perceptual processes. As mentioned, independence can dramatically reduce the perception-to-action mappings by using factorial codes rather than complex codes based on the full joint distribution. Hence, it is a natural starting point to look for high-level statistically independent features when aiming at high-level representations. In this paper we focus on cognitive processes in digital speech signals. The paper is organized as follows: First we discuss the specifics of the cognitive component hypothesis in relation to speech, then we describe our specific methods, present results obtained for the TIMIT database, and finally, we conclude and draw some perspectives.

Cognitive Component Analysis

In sensory coding it is proposed that visual system is near to optimal in representing natural scenes by invoking 'sparse distributed' coding (Field, 1994). The sparse signal consists of relatively few large magnitude samples in a background of numbers of small signals. When mixing such indepen-

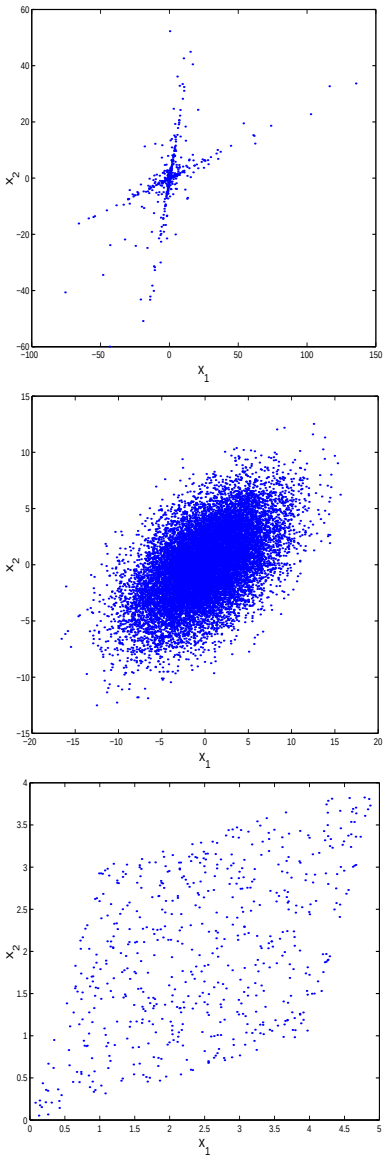

Figure 1: Prototypical feature distributions produced by a linear mixture, based on sparse (top), normal (middle), or dense source signals (bottom), respectively. The characteristics of a sparse signal is that it consists of relatively few large magnitude samples on a background of weak signals, hence, produces a characteristic ray structure in which the ray is defined by the vector of linear mixing coefficients: One for each for a sparse source.

dent sparse signals in a simple linear mixing process, we obtain the 'ray structure' which we consider emblematic for our approach, see the top panel in Figure 1. 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 sources signals. Thus, we used ICA to model the ray structure and represent semantic structure in text, social networks, and other abstract data such as music

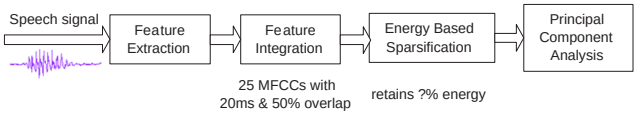

Figure 2: Preprocessing pipeline for speech COCA. MFCCs are extracted at the basic time scale (20*ms*). According to applications, features are averaged/stacked into longer time scales. Energy based sparsification is followed as a method to reduce intrinsic noise. PCA on sparsified features projects on a relevant subspace that makes it possible to visualize the 'ray'-structure. A subsequent ICA can be used to identify the actual ray coordinates and source signals.

(Hansen et al., 2005; Hansen & Feng, 2006). Within so-called bag-of-words representations of text, COCA is a generalization of principal component analysis based 'latent semantic analysis' (LSA), originally developed for information retrieval on text (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). *The key observation is that by using ICA, rather than PCA, we are not restricted to orthogonal basis vectors*. Hence, in ICA based latent semantic analysis topic vocabularies can have large overlaps. We envision that these implemented by overlapping receptive fields can detect more subtle differences than 'orthogonal' receptive fields.

Here we are going to elaborate on our earlier findings related to speech. The basic preprocessing pipeline for COCA of speech is shown in Figure 2. First, basic features are extracted from a digital speech signal 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 our frequency resolution is better at lower frequencies. These so-called mel-frequency cepstral coefficients[1] (MFCC) features are next aggregated in time. Simple energy based filtering leads to sparse representations. Sparsification is regarded as a simple means to emulate a saliency based attention process.

We have earlier reported our preliminary findings of ICA ray structure related to phonemes and speaker identity in a relatively small database (Feng & Hansen, 2005, 2006). Figure 3 illustrates the phoneme relevant ray structure at the basic time scale. This analysis was carried out on four simple utterances: 's', 'o', 'f' and 'a'. As shown in the figure, cognitive components of /e/ phoneme opening 's' and 'f' are identified.

We speculate that these phoneme-relevant cognitive components contribute towards the well-known basic invariant 'cue' characteristics of speech (Blumstein & Stevens, 1979). The theory of acoustic invariants points out that the perceived signals are derived as stable phonetic features despite of the different acoustic properties produced by different speakers. Moreover Damper has shown that although the speech signal may vary due to coarticulation, the relation between key fea-

[1]For a complete description of MFCC and related cepstral coefficients, see (Deller, Hansen, & Proakis, 2000).

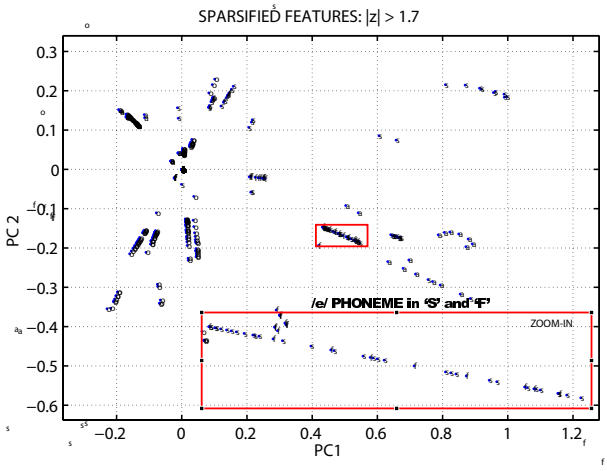

Figure 3: The latent space is formed by the two first principal components of data consisting of four separate utterances representing the sounds 's', 'o', 'f', 'a'. The structure clearly shows the sparse component mixture, with 'rays' emanating from the origin (0,0). The ray embraced in a rectangle contains a mixture of 's' and 'f' features, a cognitive component associated with the vowel /e/ sound.

tures follows a consistent and invariant form (Damper, 1998). Experiments involving labels related to speaker identification also provided the signature of linear 'ray'-structures. Is linearity related to perceptually distinguishable categories? The discussion on linear correlations in the speech signal and locus equation is still on-going (Sussman, Fruchter, Hillbert, & Sirosh, 1998).

During the itinerary of searching for spoken cognitive components, we have thus already reported (Feng & Hansen, 2005, 2006) on generalizable phoneme relevant components at a time scale of $20 \sim 40ms$, and generalizable speaker specific components at an intermediate time scale of $1000ms$.

In this paper we will further expand on our findings in speech by applying COCA on speech features at various time scales. We will systematically investigate the performance of unsupervised and supervised learning and test whether the tasks are learned in equivalent representations, hence, indicating consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels).

Methods

Our speech analysis follows the basic preprocessing scheme shown in Figure 2.

Feature Stacking

Since speech signals are non-stationary features have to be extracted from short-time scales. A simple method to get features at longer time scales is stacking or vector 'concatenation' of signals. Figure 4 illustrates the stacking procedure used in our experiments.

1. Truncate speech signal into overlapped frames, $20ms$ long with 50% overlap;

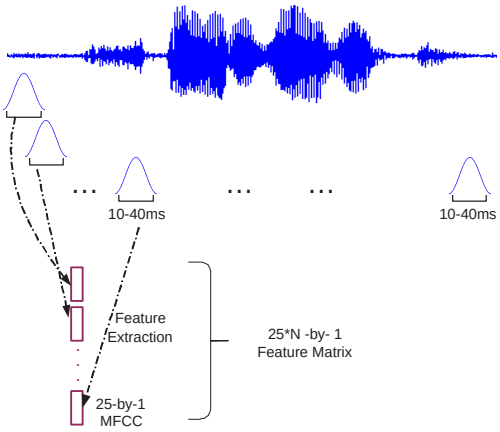

Figure 4: Speech feature extraction and stacking

2. Apply hamming window on each frame;
3. Extract MFCCs from each windowed frame, which forms a 25-dimensional vector;
4. According to the time scale, N original 25-dimensional MFCCs are stacked into one $25 * N$-dimensional vector;
5. Repeat 4 until all the frames are stacked.

$25 * N$ dimensional features representing long time scales are then used in both supervised and unsupervised learning methods.

Mixture of Factor Analyzers

To test whether supervised and unsupervised learning lead to similar representations we need a model that can incorporate both. In particular we need a generative representation to allow unsupervised learning, and we want the representation to

allow sparse linear ray like features. This can be achieved in a simple generalization of so-called mixture of factor analyzers (MFA). The unsupervised version is inspired by the so-called *Soft-LOST* (Line Orientation Separation Technique) (O'Grady & Pearlmutter, 2004).

Factor analysis is one of the basic dimensionality reduction forms. It models the covariance structure of multi-dimensional data by expressing the correlations in lower dimensional latent subspace, mathematical expression is

$$\mathbf{x} = \Lambda \mathbf{z} + \mathbf{u}, \tag{1}$$

where $\mathbf{x}$ is the p-dimensional observation; Λ is the factor loading matrix; $\mathbf{z}$ is the k-dimensional hidden factor vector which is assumed Gaussian distributed, $\mathcal{N}(\mathbf{z}|0,I)$; $\mathbf{u}$ is the independent noise which is $\mathcal{N}(\mathbf{u}|0,\Psi)$, with a diagonal matrix Ψ. Given eq. (1), observations are also distributed as $\mathcal{N}(\mathbf{x}|0,\Sigma)$, with $\Sigma = \Lambda\Lambda^T + \Psi$. Factor analysis aims at estimating Λ and Ψ in order to give a good approximation of covariance structure of $\mathbf{x}$.

While the simple factor analysis model is globally linear and Gaussian, we can model non-linear non-Gaussian processes by invoking a so-called mixture of factor analyzers

$$p(\mathbf{x}) = \sum_{i=1}^{K} \int p(\mathbf{x}|i,\mathbf{z}) p(\mathbf{z}|i) p(i) d\mathbf{z}, \tag{2}$$

where $p(i)$ are mixing proportions and K is the number of factor analyzers. MFA combines factor analysis and the Gaussian mixture model, and hence can simultaneously perform clustering, and dimensionality reduction within each cluster, see (Ghahramani & Hinton, 1996) for a detailed review.

To meet our request for unsupervised learning model, MFA is modified to form an ICA-like line based density model similar to *Soft-LOST* by reducing the factor loadings to hold a single column vector, i.e., the 'ray' vector. It uses an EM procedure to identify orientations within a scatter plot: in the E-step, all observations are *soft* assigned into K clusters depending on the number of mixtures, which is represented by orientation vectors v_i, then it calculates posterior probabilities assigning data points to lines; and in M-step, covariance matrices are calculated for K clusters, and the principal eigenvectors of covariance matrices are used as new line orientations v_i^{new}, by this means it re-positions the lines to match the points assigned to them. Finally we end up with a mixture of lines which can be used as a classifier. We purposed a supervised mode of the modified MFA, which models the joint distribution of features set $\mathbf{x}$ and a possible labels set $\mathbf{y}$

$$p(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{K} \int p(\mathbf{x}|i,\mathbf{z}) p(\mathbf{z}) d\mathbf{z} p(\mathbf{y}|i) p(i). \tag{3}$$

In the sequel we will compare the performance of the two modes of modified MFA at multiple time scales. In particular we will train supervised and unsupervised models on the same feature set. For the unsupervised model we first train using only the features $\mathbf{x}$. When the density model is optimal we clamp the mixture density model and train only the cluster tables $p(\mathbf{y}|i)$, $i = 1,...,K$, using the training set labels. This is also referred to as unsupervised-then-supervised learning. 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'.

Results

In this section we will present experimental results of analysis on speech signals gathered from TIMIT database (Garofolo et al., 1993). TIMIT is a reading speech corpus designed for the acquisition of acoustic-phonetic knowledge and for automatic speech recognition systems. It contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from the United States. For each utterance we have several labels that we think as cognitive indicators, labels that humans can infer given sufficient among of data. While each sentence lasts approximately $3s$ we will investigate performance at time scales ranging from basic $20ms$ to long about $1000ms$. The cognitive labels we will focus on here are phonemes, gender, height and speaker identity. Training and test sets are recommended in TIMIT, which contain 462 speakers reading for training and 168 for test. The total speech covers 59 phonemes, and the heights from all speakers range from $4'9''$ to $6'8''$, and have totally 22 different values. In order to gather sufficient amount of speech signals we chose 46 speakers with equal gender distribution, and speech signals cover all 59 phonemes, and all 22 heights.

Following the preprocessing pipeline, we first extracted 25-dimensional MFCCs from original digital speech signals. To investigate various time scales, we stacked basic features into a variety of time scales, from the basic $20ms$ scale up to $1100ms$. Energy based sparsification was used afterwards as a means to reduce the intrinsic noise and to obtain sparse signals. Sparsification is done by thresholding the amplitude of stacked MFCC coefficients, and only coefficients with super threshold energy were retained. By adjusting the threshold, we examine the role of sparsification in our experiments. We changed the threshold leading to a retained energy from 100% to 41%. Unsupervised and supervised modes of MFA were then performed respectively. To classify a new datum point x_{new} we first calculate the set of $p(i|\mathbf{x}_{new})$'s and then compute the posterior label probability.

Figure 5 presents the results of MFA for gender detection. The two plots (a) and (b) show the error rates for the supervised mode of MFA for the training and test set separately, while (c) and (d) are training and test error rates for unsupervised MFA (*soft-LOST*). First, we note that sparsification does play a role: when high percentage of features was retained from sparsification, e.g. 100% and 99.8%, error rates did not change much while increasing time scales, meaning the intrinsic noise covers up the informative part, and longer time scales do not assist to recover it. With the increasing of time scales all the curves tend to converge at the time scale around $400 \sim 500ms$.

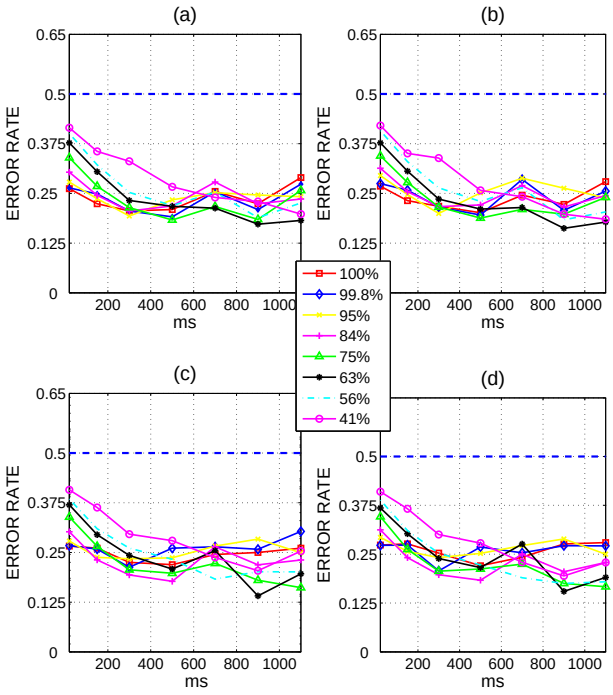

Figure 5: Error rates as function of time scales for different thresholds in gender detection. (a), (b): Training error rates and test error rates of supervised MFA respectively; (c), (d): Training error rates and test error rates of unsupervised MFA; The 8 curves represent feature sparsification with retained energy from 100% to 41%. The dashed lines are the baseline error rates for random guessing. Results indicate that the relevant time scale is about $400 \sim 500ms$ for this task.

Table 1: Timescales recommended for modeling Phonemes, Gender, Height, Identity

(*ms*)	**Phoneme**	**Gender**	**Height**	**ID**
Timescale	20	400-500	≥ 1000	≥ 1000

Similar experiments have been performed on phoneme, height and speaker identity. For phoneme recognition, the 59 phonemes from TIMIT database include vowels, fricatives, stops, affricates, nasals, semivowels and glides. To simplify the problem, we grouped these phonemes into 3 large categories: Vowels, fricatives and others. Stacking features into longer time scales for phoneme recognition degrades the performance, which shows consistency with our previous work that phonemes are best modeled at short time scale. The results of all experiments are summarized in Table 1.

To illustrate how well supervised and unsupervised representations are aligned, we follow the approach outlined above. We trained with appropriate labels in supervised mode to represent the human observer, and with the unsupervised-then-supervised scheme to represent the 'ecological' grouping. In both cases we can measure the test performance of the resulting classifier. High correlation between the error rates of the two schemes indicates similarity of the representations.

Figure 6 presents the correlation of test performance for supervised and unsupervised learning modes of MFA. For all the four classification tasks, for the given time scales and thresholds, data show a remarkable correlation. Hence, in line with the cognitive component hypothesis the statistical regularities captured by unsupervised learning are highly compatible with the cognitive structure represented by the label structures.

Conclusion

Cognitive component analysis of speech have revealed statistical regularities at multiple time scales corresponding to phoneme, gender, height and speaker identity.

We have devised a protocol for testing the cognitive component hypothesis based. We propose to compare the performance of supervised learning and unsupervised learning under closely matched conditions, so that the only difference is that 'cognitive labels' are used for supervised learning while not for unsupervised learning.

We preprocessed speech in a pipeline starting from the basic features: short time (20*ms*) 25-dimensional Mel-frequency Cepstral Coefficients (MFCCs). Feature stacking was used to aggregate features at multiple time scales. Energy based sparsification was invoked to obtain a sparse distributed representation and for noise reduction. We found that

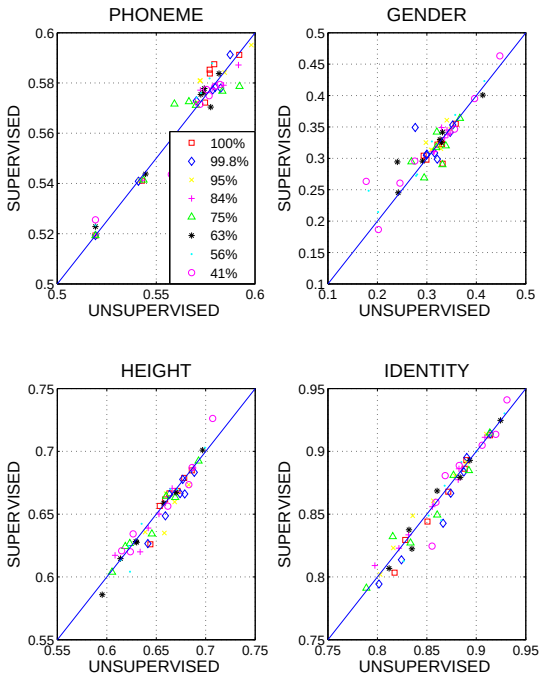

Figure 6: Correlation between test error rates of supervised and unsupervised learning on four label sets: phoneme, gender, height and identity. Solid lines indicate $y = x$ in the given coordinate systems. All data locate along this line. We can conclude that high correlation between supervised and unsupervised learning has been found for a wide variety of error rates substantiating our claim that two representations are highly similar.

the following time scales are characteristic: $20ms$ of speech provides phonemes information; gender is found in the range $400 \sim 500ms$; while, height and identity may require longer time scales, say $> 1000ms$.

Our finding indeed indicates the consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels), for phonemes, gender, speaker identity all of which are effortless recognized by humans. Height is also predicted from speech features corresponding to human ability to guess the speakers size. It would be interesting to test whether our representations lead to similar errors in predicting a persons height from speech as in humans.

Acknowledgments

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound' (STVF No. 26-04-0092), www.intelligentsound.org. We thank Tobias Andersen for useful comments on the manuscript. LF thanks the Niels Bohr Legatet for generous financial support for external research stay.

References

Barlow, H. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*, 1001–1017.

Damper, R. I. (1998). Self-learning and self-organization as tools for speech research. *Behavioral and brain sciences*, *21*, 262–263.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Deller, J. R., Hansen, J. H., & Proakis, J. G. (2000). *Discrete time processing of speech signals*. IEEE Press Marketing.

Feng, L., & Hansen, L. K. (2005). On low level cognitive components of speech. In *Proc. international conference on computational intelligence for modelling* (Vol. 2, pp. 852–857).

Feng, L., & Hansen, L. K. (2006). Phonemes as short time cognitive components. In *Proc. icassp* (Vol. 5, p. 869-872).

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). The darpa timit acoustic phonetic continuous speech corpus cdrom. In *Nist order number pb91-100354*.

Ghahramani, Z., & Hinton, G. E. (1996). *The em algorithm for mixtures of factor analyzers* (Tech. Rep. No. CRG-TR-96-1). 6 King's College Road, Toronto, Canada M5S 1A4: University of Toronto, Department of Computer Science.

Hansen, L. K., Ahrendt, P., & Larsen, J. (2005). Towards cognitive component analysis. In *Akrr'05 -international and interdisciplinary conference on adaptive knowledge representation and reasoning*.

Hansen, L. K., & Feng, L. (2006). Cogito componentiter ergo sum. In *Proc. ica* (pp. 446–453).

Hoyer, P., & Hyvrinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Comput. Neural Syst.*, *11*, 191–210.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*, 356–363.

O'Grady, P. D., & Pearlmutter, B. A. (2004). Soft-lost: Em on a mixture of oriented lines. In *Proc. ica* (p. 430-436).

Sussman, H. M., Fruchter, D., Hillbert, J., & Sirosh, J. (1998). Linear correlations in the speech signal: The orderly output constraint. *Behavioral and brain sciences*, *21*, 241–299.

Wagensberg, J. (2000). Complexity versus uncertainty: The question of staying alife. *Biology and philosophy*, *15*, 493–508.