On Low-level Cognitive Components of Speech

Ling Feng
Intelligent Signal Processing,
Informatics and Mathematical Modeling,
Technical University of Denmark, Denmark
lf@imm.dtu.dk

Lars Kai Hansen
Intelligent Signal Processing,
Informatics and Mathematical Modeling,
Technical University of Denmark, Denmark
lkh@imm.dtu.dk

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 speakers. 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 (ICA) algorithms [1]. It is an intriguing fact that representations found in human and animal perceptual systems closely resembles the theoretically optimal representations obtained by independent component analysis on visual contrast detection [2], on visual features involved in color and stereo processing [3], and on representations of sound features [4].

Ref. [5] defined and investigated the independent *cognitive* component hypothesis, which basically asks the question: *Do humans also use these information theoretically optimal 'ICA' methods in more generic and abstract data analysis.* We proposed to use the term *cognitive component analysis* (COCA) for unsupervised learning algorithms that present such 'spontaneous cognition'.

Here we are interested in pursuing this idea in the context of speech. We are interested in purely auditory aspects, not contents *per se*. We will focus on two aspects, phoneme features and speaker features. Our presentation will be qualitative, mainly 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 *labeled* examples [6]. In this case classification belongs to the class of *supervised* learning problems. On the other hand clustering which is related to *unsupervised* learning problem, uses general statistical rules to group objects, without a priori providing a set of labeled 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. Grouping by ICA has been earlier pursued for 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 these research works that ICA is a more appropriate model than both principal component analysis (PCA), which is too constrained, and clustering, which may in some instances be too flexible as a representation of text data [5].

2. Cognitive component analysis

Lee and Seung introduced the method of non-negative matrix factorization (NMF) [12] as a scheme for parts-based object recognition. The factorization of an observation matrix in terms of a relatively small set of *cognitive components* leads to a parts-based object representation. The values of the non-negative representation for objects in images and text have been demonstrated. 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 [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) is a very useful tool for dimensionality reduction and may be 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 [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 here 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 simple Euclidean distance metric can be used if document vectors are properly normalized, otherwise angular distance may be used. This approach is principled, fast, and language independent. Deerwester and co-workers developed the concept of latent semantics based on PCA of the term-document matrix [15]. The fundamental observation behind the LSI approach is that similar documents use similar vocabularies, hence, the term 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, 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 dimensionality reduction tool, which is then post-processed to reveal cognitive components, e.g., by interactive visualization schemes [17].

2.2. Independent component analysis

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 and the sources are statistically independent processes, it is often possible to recover sources and mixing, using a variety of ICA 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 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 $\mathbf{AA}^{\mathrm{T}}$ is not enough to uniquely identify **A**, since if one solution **A** is found, any (row) rotated matrix $\widetilde{\mathbf{A}} = \mathbf{AU}$, $\mathbf{UU}^{\mathrm{T}} = \mathbf{I}$ is also a solution, because $\widetilde{\mathbf{A}}$ has the same outer product as **A**.

This is a potential problem for LSI based analysis. If the document database can be modeled as in (1) then the original characteristic context histograms will not be found by LSI. The field of ICA has on the other hand devised many algorithms that use more informed statistics to locate **A** and thus **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 Fig. 1. In the upper panel of Fig. 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 (1), we obtain a set of 'rays' emanating from the origin. The directions of the rays are given by the column vectors of the **A**-matrix. If the sources are normally distributed (middle panel of Fig. 1) there is no additional information but the covariance matrix. Hence, in some sense this is a worst case for separation. Fortunately, many interesting real world data sets are very sparse, hence, more similar to the upper panel of Fig. 1.

3. Component analysis of speech

In the authoritative textbook ‘Discrete-Time Processing of Speech Signals’ by Deller et al. [18] 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 [18].

The processes in the speech production system are generally considered stationary for time intervals on the order of 20 msec [18], 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 10 kHz, 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’ [18]. The cepstral coefficients are used in speaker independent speech recognition, because in this context the pitch is a confound. The speaker dependent and speaker independent aspect are separated in the cepstral coefficient representation, hence, we use this representation 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 [19] were used for this demonstration. For the analysis we used 20 msec analysis windows with 50% overlap. The windows were represented by 16 cepstral coefficients. The temporal development of the cepstral representation of the four utterances is presented in two versions in Fig. 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 with a statistical z < 1.7. In Fig. 3 we show the scatter plot of the set of windows projected onto the first two principal components derived from the 16 x 16 sparsified 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

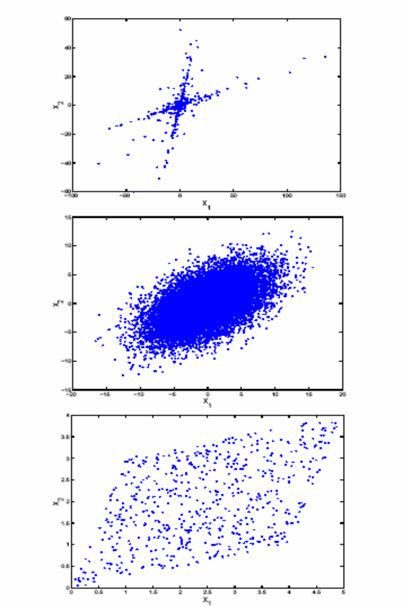

Fig. 1. Prototypical feature distributions
Prototypical feature distributions produced by a linear mixture, based on sparse (top), normal (middle), or dense source signals (bottom), respectively. The characteristic of the sparse signal is that it consists of relatively few large magnitude samples on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

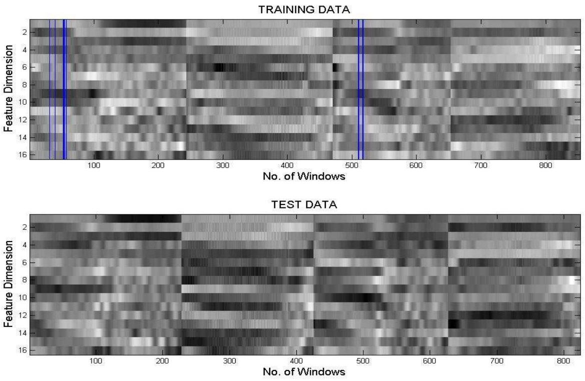

Fig. 2. Cepstral coefficient sequences for training and test sets
Four separate utterances are concatenated for this experiment, representing the sounds ‘s’, ‘o’, ‘f’, ‘a’. Each concatenated set of utterances is represented twice: in a training set and in a test set. The boundaries between the four utterances 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 of vertical lines to indicate positions of analysis windows that belong to a generalizable finger print feature further discussed in Fig. 3.

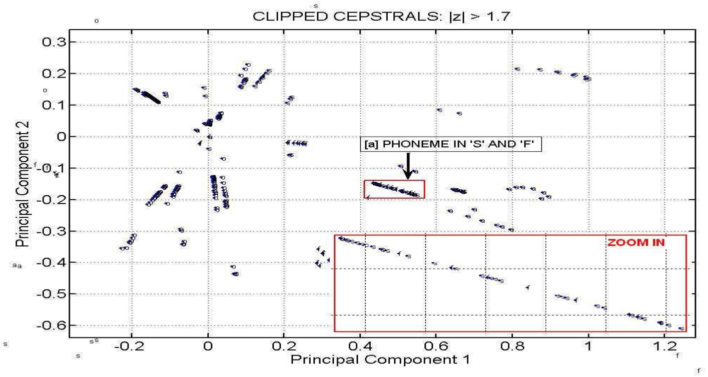

Fig. 3. Scatter plot of data on latent space

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 Fig. 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 windows were indicated by vertical lines in Fig. 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'.

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 belong to, we have marked up several points (windows) in Fig. 3 and have indicated the temporal location of these windows as vertical stripes in Fig. 2. 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 proved by creating a similar plot with the projections of the *test set* windows (data not shown). This structure is indeed generalizable in contrast to some of the other ray-like structures that apparently are too 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 phonemes are not aligned with the directions of the principal components, hence, an ICA scheme is required. Phoneme recognition is an active research field in speech recognition, see e.g., [20], 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 role both in social contexts and in speech based engineering applications, e.g., related to access control [21].

Speaker recognition has two aspects: Speaker identification, 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 [22]. 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. Speaker recognition systems are moreover 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. The most widely accepted features for speaker recognition are mel-frequency cepstral coefficients (MFCC). The MFCCs are perceptually weighted cepstral coefficients [18].

According to our basic hypothesis the speaker dependent generalizable 'cognitive' components should be elucidated by Latent Semantic Indexing (LSI). To test the hypothesis we study here three speakers' voice messages from our in-house ELSDSR speech database [23]. In this database, read text is recorded using a MARANTZ PMD670 portable solid state recorder, and stored in PCM (wav) format. The sampling frequency is 16 kHz. ELSDSR contains voice messages from a total of 22 speakers (12M/10F) 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 [24].

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 windows with 20 msec signal content were blocked without overlap, and 12 MFCCs were extracted from each window. To form the long-term features, 50 basic analysis windows were concatenated. The dimensionality of the aggregate representation is thus 50 x 12. The total number of such expanded windows in the analysis was 522. After variance normalization, energy based sparsification was performed on the high dimensional data, and the upper 1% fraction was retained. Finally, LSI (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 Fig. 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, blue and black points emanate from (0,0), and show similar sparse ICA 'ray' structures. These features of same text also carry characteristics 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 in Fig. 4. 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 data and training data, which is speaker dependent. We therefore stipulate that this effect is an interaction between the text content and the speaker identity.

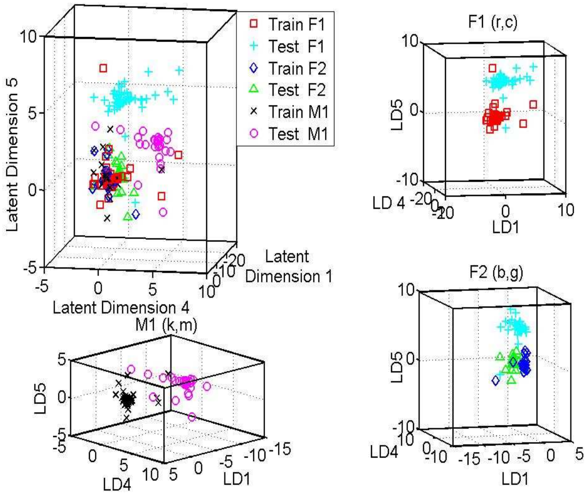

Fig. 4. Text-dependent speaker recognition
We focus on text-dependent speech. The basic analysis window of the speech signal is represented by 12 MFCCs. 50 basic analysis windows are concatenated to form an intermediate time scale representation. We sparsified the coefficients by retaining the upper 1% magnitude fraction. We used a training set from speakers F1, F2 and M1. The data from the training set is submitted for LSI, we show the scatter plots of both training and test data in the space of the 1st, 4th and 5th latent components. The upper left display shows all data points. 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 basic time scale analysis window (20 msec). The structure is indeed speaker dependent in the sense that the ray systems are offset from the origin. We conclude that we find a mixture of phoneme like features and speaker identity features.

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 2nd, 4th and 5th. Fig. 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. Thus the generalizable ray structure emanates from (0,0) *without* offsets.

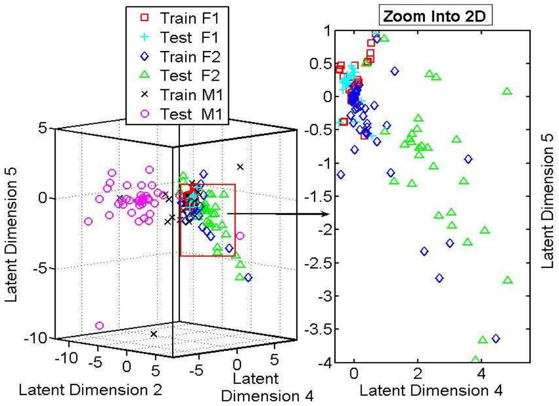

Fig. 5. Text-independent speaker recognition
We focus on text-independent speech. The setup is the same as text-dependent case. In the left panel all data points are shown as represented in the space of the 2nd, 4th and 5th latent components. There is an evident ray structure corresponding to a generative ICA model based on linear mixing of sparse sources. In contrast to the text-dependent case 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.

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 derived cognitive components of speech signals. We used homomorphic filtering to derive features, and 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 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. 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] 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] John R. Deller, John H. Hansen, and John G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press Marketing, 2000.

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

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

[21] *http://www.research.ibm.com/VIVA Demo*, 2005.

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

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

[24] 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.