Cognitive Components of Speech at Different Time Scales

Ling Feng
Informatics and Mathematical Modeling
Technical University of Denmark
B.321 Lyngby, Denmark
lf@imm.dtu.dk

Lars Kai Hansen
Informatics and Mathematical Modeling
Technical University of Denmark
B.321 Lyngby, Denmark
lkh@imm.dtu.dk

Abstract

We discuss the cognitive components of speech at different time scales. We investigate cognitive features of speech including phoneme, gender, height, speaker identity. Integration by feature stacking based on short time MFCCs. Our hypothesis is basically ecological: we assume that features that essentially independent in a reasonable ensemble can be efficiently coded using a sparse independent component representation. This means that supervised and unsupervised learning should result in similar representations. We do indeed find that supervised and unsupervised learning of a model based on identical representations have closely corresponding abilities as classifiers.

1 Introduction

We are interested in modelling medium to high level human representations of sound. Human perception and cognition has evolved through a long-time adaptation process in the face of natural environment statistics. We envision that efficient representations of high level processes are based on sparse codes and approximate independence as has been found for more basic perceptual processes. To nail down the nature of these representations we follow the approach that has been invoked for modelling low level perception namely to study natural ensemble statistics, see e.g., [1, 2]. Obviously, robust statistical 'regularities' will be exploited by an evolutionary optimized brain [3]. Statistical independence may be one such regularity. Independence can dramatically reduce the complexity of perception-to-action mappings, hence, it is a natural starting point to look for high-level statistically independent features of speech when aiming at high-level representations.

Human perceptual systems can model complex multi-agent scenery, and has the ability of using a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing agents. Independence is an assumption that has found significant support in the discussion of multi-agent audio, in the so-called cocktail party problem, see e.g., [4, 5]. The resulting optimized representations achieved by a variety of ICA algorithm closely resemble representations found in human perceptual systems on visual contrast detection [1], on visual features involved in color and stereo processing [6], and on representations of sound features [2].

Within an attempt to generalize these findings to higher cognitive functions, we have investigated the independent cognitive component hypothesis, which basically asks the question: *Do humans use information theoretically optimal 'ICA' methods in more generic and abstract data analysis.*

We introduced cognitive component analysis (COCA) in [7]: the process of unsupervised grouping of data such that the resulting group structure is well-aligned with that resulting from human cognitive activity. The basic scheme of COCA is shown in Fig. 1 and has produced evidence that ICA is relevant for representing semantic structure in text, social network, and other abstract data, e.g. musical features [7, 8].

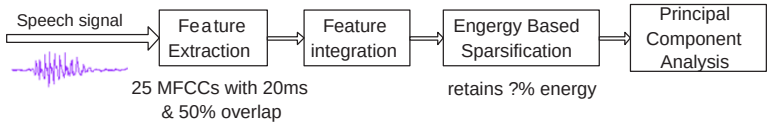

Figure 1: Preprocessing pipeline for sognitive component analysis (COCA) of speech. MFCCs are extracted at a basic time scale ($20ms$). Features are integrated in windows by temporal stacking. To reduce noise, energy based sparsification is performed, which is followed by PCA to project features into a latent semantic space.

We have approached speech with the same agenda in [9, 10]. At a basic time scale ($20 \sim 40ms$) we found generalizable phoneme relevant components while at an intermediate time scale ($< 1s$) we discovered generalizable speaker-specific sparse components. These studies used the mel frequency weighted cepstral coefficients (MFCC's) as basic low level representation.

Meng et al. [11] studied time scales in relation to cognitive levels of music. They showed short time scale ($30ms$) were relevant to concept linked to instantaneous frequency, e.g., harmonics and pitch; medium time scale ($740ms$) to timbre, modulation; and long time scales ($9.62s$) to more abstract information such as beat and 'mood'.

Here we will further expand on our findings in speech COCA. We will systematically investigate the performance of unsupervised and supervised learning to find features that are learned in equivalent representations, hence, indicating consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels).

2 Methods

Our speech analysis preprocessing pipeline is shown in Fig. 1 We use a simple 25-dimensional mel frequency weighted ceptral coefficient as short time feature (MFCC) (coefficients and temporal differences of coefficients within an analysis window of $20ms$). To explore speech at multiple time scales we stack the basic signals in windows corresponding to the time scale of interest. A simple moving average within windows gave similar results. We further reduce noise by energy based sparsification at different thresholds emulating a simple saliency based attention process as in [7].

2.1 Mixture of factor analyzers

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

While the factor analysis model is basically linear we can model non-linear manifolds by invoking a so-called mixture of factor analyzers (MFA)

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

where w_i are mixing proportions; $p(\mathbf{z}|w_i) = p(\mathbf{z})$; K is the total number of factor analyzers. MFA can be seen as a combination of FA and Gaussian mixture model, and hence can simultaneously perform clustering, and dimensionality reduction within each cluster, see [12] for a detailed review.

MFA is here modified to form an ICA-like line based density model, similar to the so-called *Soft-LOST* (Line Orientation Separation Technique) of [13]. It uses an EM procedure to identify orientations within a scatter plot: first in E-step, all the observations are *soft* assigned into S clusters

Table 1: Timescale of phoneme, gender, height, identity

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

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 S clusters, and the principal eigenvectors of covariance matrices are used as new line orientation v_i^{new}, by this means it repositions the lines to match the points assigned to them. This method is an unsupervised learning method. It can be turned into supervised learning method to model the joint distribution of features and a possible labels set $\mathbf{y}$,

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

where $\mathbf{y}$ is the label set for observations $\mathbf{x}$. In the sequel we will compare the performance of the two models at multiple time scales. In particular we will train supervised and unsupervised models on the same data set. For the unsupervised model we first train using only the features $\mathbf{x}$. When the density model is optimal we clamp the mixture model and train the cluster tables $p(\mathbf{y}|w_i)$ using the training set labels.

3 Results

In this section we will show the experimental results of COCA analysis on speech signals gathered from the TIMIT speech database. It collects 630 speakers, which covers 59 phonemes, and 22 different 'heights'. For each speaker, we have approximately $30s$ from 10 sentences.

3.1 What's between phonemes and identity?

As mentioned in Sec. 1, phonemes as the smallest linguistic units need short time scale to be recognized, on the order of $20ms$; while speaker identity requires substantially longer time integration. We hypothesize that gender needs shorter time scale than identity.

In 2003 it was found that estimating height from speech is possible [14]. Furthermore it has been proved that there is a correlation between speaker's vocal tract length and speaker's height [15]. We therefore include height estimation as a label to predict from speech. If height is a predictable quantity we can in future studies use this as a test of our approach: Can humans guess heights from speech? - and is the confusion similar to that of the machine learning algorithm?

We thus consider four different cognitive phenomena: phoneme, gender, height, and speaker identity. In order to gather speech signals which can represent sufficient information w.r.t. these topics, we chose 46 speakers with equal gender distribution, speech signals cover all the 59 phonemes, and the 22 different values of heights within TIMIT which are from 4 feet 9 inch to 6 feet 8 inch. Energy based sparsification is used as a means to reduce the intrinsic noise, and to obtain sparse sources.

We stacked features with a variety of time scales, from $20ms$ to $1100ms$, and sparsified stacked features with diverse thresholds with a retained energy ranging from 100% to 41%. Fig. 2 gives the results of MFA on gender detection. We can say that sparsification does play a role: when high percentage of features was retained, e.g. 100% and 99.8%, error rates did not change much while increasing time scales, meaning the intrinsic noise degrade the informative part, and longer time scales do not assist to recover it; when too few features survived, e.g. 58% and 41%, classification error began to grow, meaning too much valuable information has been removed. Overall retaining about 75% of the energy provides the best performance, and when the time scale is around $400 \sim 500ms$, minimal error rate is obtained.

Similar experiments have been performed on phoneme, height and identity. The results are summarized in Table 1.

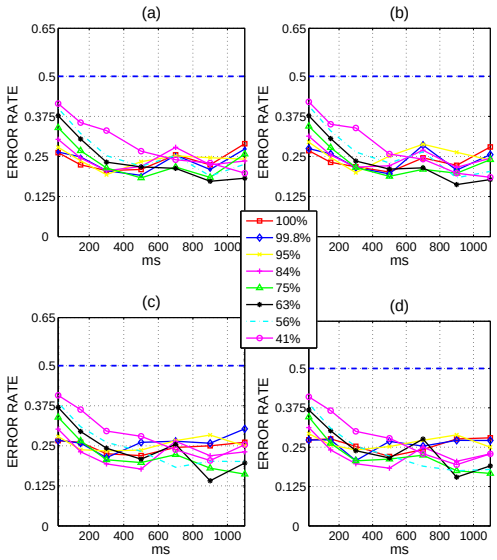

Figure 2: The figure shows the error rates with increasing time scales and thresholds for gender detection. (a),(b): Training and test error rates of supervised learning respectively; (c),(d): Training and test error rates of unsupervised learning. 8 curves show feature sparsification with retained energy from 100% to 41%. Dash lines are the baseline for random guessing. Results indicate the $400 \sim 500ms$ time scale for gender.

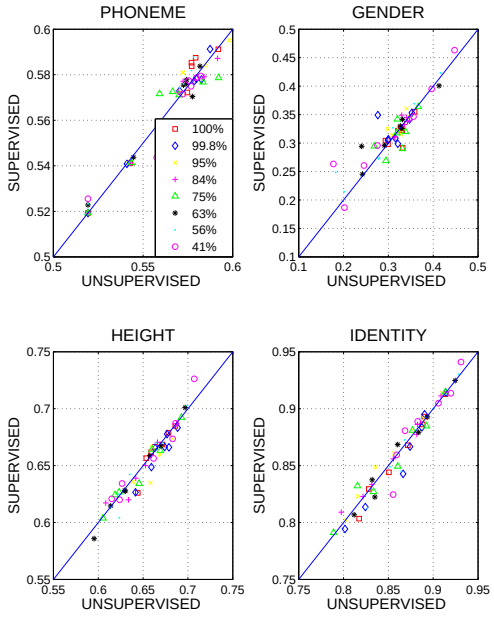

Figure 3: Correlation between test error rates of supervised and unsupervised learning on the four label sets: phoneme, gender, height and identity. Solid lines indicate $y = x$ in the given coordinate systems. We find very high correlation between supervised and unsupervised learning for a wide variety of error rates substantiating our claim that the two representations are highly similar.

3.2 Unsupervised vs. supervised learning

To illustrate how well supervised and unsupervised representations are aligned, we follow the approach outlined above. We trained with appropriate labels in supervised mode and with the unsupervised-then-supervised scheme. 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.

Fig. 3 presents the correlation of test performance for supervised and unsupervised learning. For all the four classification tasks at given time scales and thresholds, data show a remarkable correlation. Hence, the statistical regularities captured by unsupervised learning is highly compatible with the cognitive structure represented by the label structures. This holds for the most obvious cognitive labels: Phoneme, gender and identity, and also for the less obvious variable 'height'.

4 Conclusion

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

We analyze speech in a pipeline starting from a basic 25-dimensional short time ($20ms$) mel frequency weighted cepstral coefficient representation. Feature stacking was used to integrate features at multiple time scales. Energy based sparsification was invoked for noise reduction.

Our results show that the following time scales are involved: $20ms$ of speech provides phoneme information; gender is found in the range $400 \sim 500ms$; furthermore, height and identity require time scale $> 1000ms$.

Acknowledgments

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound' (FTP No. 26-04-0092), www.intelligentsound.org. LF thanks the Otto Mønsted Fond for generous financial support for the external research visit.

References

[1] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, pp. 3327–3338, 1997.

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

[3] H.B. Barlow, "Unsupervised learning," *Neural Comp.*, vol. 1, 295-311, 1989

[4] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comp.*, vol. 17, 1875–1902, 2005.

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

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

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

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

[9] L. Feng and L. K. Hansen, "On low level cognitive components of speech," in *Proc. International Conference on Computational Intelligence for Modelling*, vol. 2, pp. 852–857, 2002.

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

[11] A. Meng, P. Ahrendt, and J. Larsen, "Improving music genre classification by short-time feature integration," in *Proc. ICASSP*, vol. 5, pp. 497–500, 2005.

[12] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," in *Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto*, 1996.

[13] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. ICA*, pp. 430–436, 2004.

[14] J. Gonzalez, "Estimation of speakers' weight and height from speech: A re-analysis of data from multiple studies by lass and colleagues," *Percept Mot Skill*, pp. 297–304, 2003.

[15] S. Dusan, "Estimation of speaker's height and vocal tract length from speech signal," in *Proc. INTERSPEECH*, pp. 1989–1992, 2005.