Data-driven Speech Representations for NMF-based Word Learning

Joris Driesen, Jort F. Gemmeke, Hugo Van hamme

Dept. ESAT, KU Leuven, Leuven, Belgium

(joris.driesen|jort.gemmeke|hugo.vanhamme)@esat.kuleuven.be

Abstract

State-of-the-art solutions in ASR often rely on large amounts of expert prior knowledge, which is undesirable in some applications. In this paper, we consider a NMF-based framework that learns a small vocabulary of words directly from input data, without prior knowledge such as phone sets and dictionaries. In the context of this learning scheme, we compare several spectral representations of speech. Where necessary, we propose changes to their derivation to avoid the usage of prior linguistic knowledge. Also, in a comparison of several acoustic modelling techniques, we determine what model properties are beneficial to the framework's performance.

Index Terms: keyword learning, non-negative matrix factorisation, clustering, acoustic modelling

1. Introduction

After decades of progress, Automatic Speech Recognition (ASR) has improved to the point of recognising huge vocabularies of words with accuracies that are not so far from those of humans, especially under favourable conditions. A downside, however, is the fact that ASR relies heavily on prior linguistic information such as phone sets describing the sounds of speech, dictionaries describing all the words in the vocabulary, context-dependency trees, co-articulation rules, etc. All this knowledge must be entered into the system by human experts, which is a time-consuming and expensive process. Moreover, acoustic models are typically trained using tens to hundreds or even thousands of hours of speech.

This reduces the flexibility of ASR, because much of this prior knowledge is language- or dialect-specific, and acoustic models only generalise well to speakers with the roughly the same speech characteristics. In contrast, we consider the group of people who suffer from a degenerative illness which affects the functioning of their upper limbs. Their pathology may affect their speech patterns in ways that are not covered by any standard acoustic or linguistic model, and vary from speaker to speaker. For them to operate assistive devices by voice ASR systems are needed which can discover relevant speech-related information with a minimum of prior knowledge, based on inputs provided by the end user [1]. Also in areas like robotics there is a growing interest in such "self-learning voice interfaces" [2, 3].

The automatic discovery of linguistic information has received considerable attention in recent years. For instance, a variety of methods for data-driven discovery of acoustic units have been proposed, e.g. [4, 5, 6], as well as methods to discover entire words or word-like patterns in acoustic data, e.g. [7, 8]. For a more thorough discussion of related methods, we refer the reader to [9] and the references therein. As a first step toward more complex systems, in this paper we consider the task of speaker-dependent word finding - building acoustic models of a select set of keywords from utterances containing multiple words, without knowing their order or location in the sentence. In [10], a word learning (keyword acquisition) method based on Non-negative Matrix Factorisation (NMF) was proposed. In the NMF-based approach, sentences are represented as a single vector indicating the presence of sound events or sequences of sound events, and each sentence is associated with one or multiple keyword labels. Word learning is done by factorising the collection of sentence-level observations into a matrix describing the features of individual keywords, and a matrix indicating the presence of these keywords in the observed utterances.

The sentence-level feature vector is created by quantising a spectral representation into a sequence of sound events, and then converting it into a histogram of sound events or histogram of co-occurring sound events (HAC) [10]. As such, the effectiveness of the word learning relies critically on the *spectral representation* that is used, and the method of quantising the spectral representation into a *intermediate representation*. Our contribution in this work is threefold. First, we investigate to what extent the effectiveness of NMF-based word learning can be improved by using discriminative features employed in modern ASR systems, instead of MFCC features [11]. In this work, we will make use of the Mutual Information Discriminant Analysis (MIDA) features proposed in [12]. Second, since discriminative features traditionally employ a set of predefined phone classes which constitute exactly the sort of prior knowledge we wish to avoid, we will propose a data-driven approach to generate MIDA features. Finally, we will compare the effectiveness of several methods to quantise the spectral representation into an intermediate representation of sound

events, based on K-means clustering and Gaussian Mixture Models (GMMs) [10, 13].

The rest of the paper is organised as follows: in section 2 we describe the NMF-based method for word learning. The spectral representations considered in this paper are described in section 3 and the intermediate representations in 4. In section 5, we describe the experimental setup, such as the keyword acquisition task used for evaluation, the results of which are shown and discussed in section 6. The conclusion and presentation of future work follows in section 7.

2. A Computational Framework for Word Learning

2.1. Non-negative Matrix Factorisation (NMF)

NMF is a technique to decompose a non-negative matrix $\boldsymbol{V}$ of size $M \times N$ into a product of non-negative matrix factors $\boldsymbol{W}$ and $\boldsymbol{H}$ of respective sizes $M \times R$ and $R \times N$ and $R \ll M$ and $R \ll N$. We write: $\boldsymbol{V} \approx \boldsymbol{W} \cdot \boldsymbol{H}$. This factorisation is solved by minimising the Kullback-Leibler divergence

$$D_{KL}(\boldsymbol{V}||\boldsymbol{W}\boldsymbol{H}) = \sum_{ij} V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \tag{1}$$

which is done by alternatingly applying multiplicative updates to $\boldsymbol{W}$ and $\boldsymbol{H}$ [14, 15]. After factorisation, columns of the matrix $\boldsymbol{H}$ indicate, for each column in $\boldsymbol{V}$ (representing an utterance), which patterns (columns in $\boldsymbol{W}$) are present. Formally, to learn words with NMF we need an operator $\psi(\cdot)$, which converts variable-length speech segments into non-negative vectors of a fixed dimensionality M. This operator must be such that for any utterance U, consisting of the words $(w_1, w_2, \ldots, w_p)$, holds:

$$\psi(U) = \psi(w_1) + \psi(w_2) + \ldots + \psi(w_p) \tag{2}$$

Applying $\psi(\cdot)$ on a set of N different speech utterances allows the creation of the N columns in a data matrix $\boldsymbol{V}$. After decomposing $\boldsymbol{V}$ with NMF, the R columns of $\boldsymbol{W}$ in principle contain representations $\psi(w_i)$, for all different words in the data.

2.2. Weakly Supervised Training

In order to learn the association between the word representations $\psi(w_i)$ in $\boldsymbol{W}$ and the keyword labels provided with the observed speech segments, we add supervision information to $\boldsymbol{V}$. The supervision also helps to improve the convergence of the NMF-based representations to keyword representations and to avoid local optima. Renaming $\boldsymbol{V}$ and $\boldsymbol{W}$ to $\boldsymbol{V_1}$ and $\boldsymbol{W_1}$ respectively, we rewrite the factorisation as

$$\begin{bmatrix} \boldsymbol{V_0} \\ \boldsymbol{V_1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W_0} \\ \boldsymbol{W_1} \end{bmatrix} \boldsymbol{H} \tag{3}$$

Where $\boldsymbol{V_0}$ marks which keywords are contained in each of the utterance representations in $\boldsymbol{V_1}$:

$$V_{0,ij} = \begin{cases} 1 & \text{if word } i \text{ in utterance } j \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

The matrix $\boldsymbol{W_0}$ is initialised as

$$\boldsymbol{W_0} = [\boldsymbol{I_K}|\boldsymbol{G}] \tag{5}$$

with K being the number of keywords, $\boldsymbol{I_K}$ a $K \times K$ identity matrix with random positive values of magnitude $O(1e-4)$ added to it, and $\boldsymbol{G}$ a $K \times (R-K)$ matrix with random positive values of magnitude $O(1e-4)$. Note that $\boldsymbol{I_K}$ is allowed to update even though in practise it does not diverge much from the identity matrix. This setup leads to solutions where each of the K keywords is mostly assigned to a single column in $\boldsymbol{W}$. The remaining $(R-K)$ columns typically model all non-keyword (filler) input.

2.3. Evaluation

In a set of previously unseen testing utterances, converted with $\psi(\cdot)$ into a matrix $\boldsymbol{V_1^{(tst)}}$, the K keywords are detected by using their representations discovered during training, i.e. $\boldsymbol{W_1}$. Concretely, we solve

$$\boldsymbol{H^{(tst),*}} = \arg\max_{\boldsymbol{H^{(tst)}}} D_{KL}(\boldsymbol{V_1^{(tst)}}||\boldsymbol{W_1}\boldsymbol{H^{(tst)}}) \tag{6}$$

after which the activation matrix $\boldsymbol{A}$ is calculated:

$$\boldsymbol{A} = \boldsymbol{W_0} \cdot \boldsymbol{H^{(tst),*}} \tag{7}$$

The matrix $\boldsymbol{A}$ is a prediction of $\boldsymbol{V_0^{(tst)}}$, the unobserved part in the testing data. The accuracy is determined as the Unordered Error Rate (UER), obtained by comparing the n_j keywords present in each utterance j with the keywords indicated by the n_j highest values in the corresponding column of $\boldsymbol{A}$. Formally:

$$UER = 100 \frac{\sum_{j=1}^{N} \#substitutions_j}{\sum_{j=1}^{N} n_j} \% \tag{8}$$

Note that the UER can only be calculated if the same keyword does not occur twice in a testing utterance and if the number of keywords occurring in each testing utterance is known. While this would not be the case most realistic applications of this learning framework, research in [9] has shown it is a good measure of the system's accuracy.

3. Spectral Representations

The basis for the majority of speech processing applications is a framing and windowing of the acoustic signal followed by the application of a Mel-scale filterbank, a set of bandpass filters whose bandwidth and spacing are based on human auditory perception. This filtering gives

rise to log-Mel spectra [11]. In this paper, frames of 25ms are shifted over the signal in increments of 10ms and weighted with a Hamming window. Applying a Mel filterbank results in a total of 22 Mel-spectral coefficients in each frame. Since the Mel-spectral coefficients are correlated, they are sub-optimal for further modelling of speech. In part, this is because the correlations in this representation introduce redundancy which causes its dimensionality to be unnecessarily high. Also, the correlations themselves are difficult to capture with computationally efficient models, popular in ASR, such as diagonal covariance GMMs. Decorrelation and dimensionality reduction is also important in the NMF-based word learning framework, since models such as GMMs are used to create intermediate representations (c.f. Section 4). We compare two methods to enhance the spectral representations of speech: MFCC features [11] and discriminative features called Mutual Information Discriminant Analysis (MIDA) features [12]. In addition, we propose a data-driven method to derive MIDA features that does not depend on prior knowledge such as a phone set.

3.1. MFCC features

MFCCs are obtained by applying an Inverse Discrete Cosine Transform (IDCT) to log-Mel spectra and thus describe the shape of these spectra in terms of high- and low-frequency cosine functions. Only a few of these coefficients, those that correspond to low quefrency components, are relevant for making phonetic distinctions between sounds. As such, they form a good low-dimensional description of the speech frame. In addition, the IDCT-transformation applied to speech shows similarities with Principal Component Analysis (PCA), a method for decorrelating features, which implies that the correlations between MFCCs are reduced compared to Mel-spectral features [16, 17, 18, 19].

In this paper, we determine 11 MFCCs in each frame, in addition to the log energy. The resulting 12-dimensional representations are then augmented with their first and second order differences (Δ- and $\Delta\Delta$-features), yielding a total of 36 coefficients per frame.

3.2. MIDA features

MIDA features are obtained with a linear transformation that maximises the separability between different classes of input frames. In this paper, we determine Δ- and $\Delta\Delta$-features on the 22 log-Mel spectral features, leading to 66-dimensional input vectors. On these representations we then perform the MIDA-transformation, separating the classes in the input space and at the same time reducing its dimensionality from 66 to 36. We used 36 resulting dimensions to keep correspondence with the dimensionality of the MFCC features). Note that this procedure differs from the creation of MFCC features described above, where dynamic information was only added at the very end.

In essence, the MIDA transformation is the concatenation of two different transformations. The first reduces the dimensionality of the input space in such a way that the loss of class information is minimal. The second transformation is performed in the reduced domain and minimises the off-diagonal values of the classes' covariance matrices. For further details, see [12]. In order to find the MIDA transformation, a frame-level classification of the training data is needed. In this paper, we use 123 speech classes, consisting of 41 phones described with 3 states. The frame-level classification is done by a forced alignment with the canonical transcription of the training data using a trained speech recogniser operating on MFCC features.

3.2.1. Data-driven MIDA features

The acoustic model used to classify the frames of the training data includes a predefined set of HMM-states with probabilistic models for their emissions, a predefined set of phones with their corresponding state sequences, and a predefined way of concatenating such phones into words. In order to avoid such use of prior knowledge, we have created a frame classification using Vector Quantisation (VQ)[20]. Our implementation of VQ relies on first applying K-means clustering the frames of the training data into $N_c = 100$ clusters, using a Euclidean distance measure operating on MFCC features. Every frame of the training data is then assigned to the cluster centre with the smallest Euclidean distance. Rather then using the entire training set however, we use a random subset of the training data, taking care that each speaker in the training data is equally represented. The resulting subset consists of 53550 frames. The MIDA transformation that optimally separates the resulting 100 data classes converts the log-Mel spectra into 36-dimensional features that we dub "VQ-MIDA" features. We will refer to the original MIDA-features obtained using an a priori defined acoustic model as "Oracle-MIDA" features.

4. Intermediate Representations

The NMF-based word learning framework used in this work does not operate directly on the spectral representations of utterances. For one, these spectral representations typically contain negative values, making them unsuitable as input for NMF. Another issue is that their dimensionality varies with the utterances' duration. Most importantly, however, they do not contain the latent linear structure put forward in (2) that allows the discovery of words. Therefore, the spectral representations are converted into an intermediate data representation with an operator $\psi(\cdot)$. In this work, this operation consists of defining a set of possible acoustic events, detecting their occurrences in the utterance, and accumulating these occurrence counts over the utterance, thus creating a histogram.

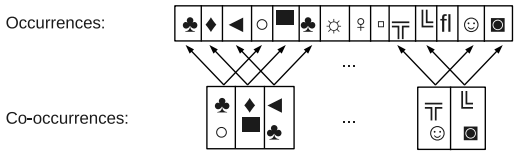

Figure 1: *Determining the combined occurrences of acoustic events at a time offset τ*

All information regarding timing and sequencing of events in the utterance is lost in this operation. To retain some of this information, we define a histogram in which *combined occurrences* of acoustic events at a certain time offset τ are counted, raising the dimensionality of the resulting histograms to the square of the number of acoustic events defined. This is illustrated in figure 1. This representation is named the *Histogram of Acoustic Co-occurrences* (HAC) [10]. In this paper, we create three HAC-vectors for each utterance, using time offsets 20ms, 50ms and 90ms, and concatenate them into a single intermediate representation. In the remainder of this section, we discuss several methods to define acoustic events based on the spectral representations described in section 3.

4.1. Vector Quantisation (VQ)

In the original proposal of NMF-based word discovery in [10], spectral features are quantised using Vector Quantisation. In this approach, each frame of a speech segment is associated with a single index, the VQ-label, and the occurrence of VQ-labels is treated as acoustic events on which HAC features are created. Using the procedure described in section 3.2.1, we create a VQ codebook of size N_c on a random subset of the training data, taking care that each speaker in the training data is equally represented. The resulting subset consists of 267750 frames. This subset is larger than the subset used to train the VQ codebook in section 3.2.1 in order to keep correspondence with the creation of GMMs described below, but pilot experiments revealed that the resulting codebooks are comparable. We refer to the use of VQ in an intermediate representation as "VQ-HAC". Concatenating the histograms for different τ-values leads to intermediate representations of dimensionality $3 \cdot N_c^2$.

4.2. Soft-VQ

As a straightforward extension to VQ, one can fit the data in each cluster with a probabilistic model [21]. In this paper, we model each VQ cluster with a *full-covariance Gaussian*. In order to avoid data scarcity, we first assign all the frames in the training data to the clusters obtained with K-means clustering, after which the covariances are obtained on the resulting class subsets. The use of a probabilistic model allows the calculation of $p(c|x_t)$ with $1 \leq c \leq N_c$, i.e. the posterior probability over the classes, given each datapoint x_t. Rather than having a binary co-occurrence of VQ-labels, we can now define the co-occurrence of posteriors as $p(c|x_t) \cdot p(c|x_{(t+\tau)})$ [13]. However, since the use of sparse co-occurrence vectors is computationally efficient, we will retain only the 3 most likely classes in each frame. We will refer to this intermediate representation as "SVQ-HAC" (Soft VQ-HAC).

4.2.1. Gaussian Mixture Models (GMM)

K-means clustering, which lies at the basis of VQ-labelling, can only discover classes that are roughly spherical, and then only if the algorithm has been initialised properly [22]. A GMM is a probabilistic model which consists of a weighted combination of Gaussians. Since a GMM can approximate any probability distribution [23, 24], they can be used to model classes of any shape, making them in theory superior to the use of VQ-labels. In this paper, shared-Gaussian GMMs are determined with the following procedure:

1. K-means clustering of the data into N_g clusters
2. Fit a diagonal-covariance Gaussian on each cluster: initial mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$ for $1 \leq k \leq N_g$
3. Initialise N_m GMMs by using agglomerative clustering on the N_g Gaussians with the KL-divergence as distance metric [9]
4. Perform EM training to update the mixture weights, means and covariances

The set of N_m GMMs (using $N_g = 1000$ shared Gaussians) thus defined is then used to generate a posterior probability for each class and each frame in a speech segment. As with VQ/SVQ-HAC, we investigate two representations: using a single label and three labels per frame, dubbed "GMM-HAC" and "SGMM-HAC" (Soft GMM-HAC), respectively. The resulting feature vectors have a dimensionality of $3 \cdot N_m^2$.

5. Keyword Acquisition task

The data on which we evaluate this framework was recorded in the context of the ACORNS project (**A**cquisition of **CO**mmunication and **R**ecog**N**ition **S**kills) [25, 26]. It consists of a total of 13188 grammatically simple English sentences in the trend of "Do you see daddy and the red ball?". These sentences contain a total number of 50 different keywords, up to 4 per sentence. They are uttered by 10 different speakers, 6 of which are male, while 4 are female. The data is split up into a training set containing 9888 randomly selected utterances, and a testing set containing the 3300 remaining ones. Each speaker is equally well-represented in both training and testing set, to reflect the fact that the purpose of NMF-based word learning is to make speaker-dependent models from user data.

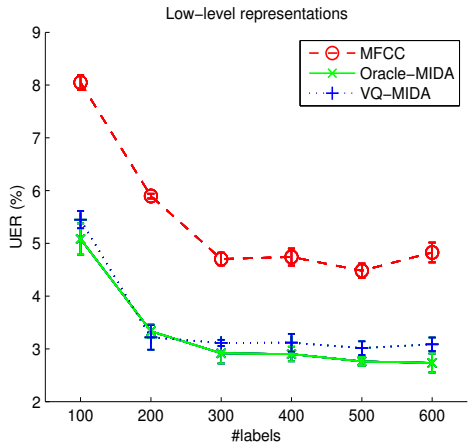

Figure 2: *Unorderered error rates (UER) obtained with different spectral representations, using VQ-HAC as intermediate representation. The horizontal axis (#labels) represents the number of clusters N_c. The vertical bars around the data points denote the standard deviation over* 5 *different random initialisations of the NMF.*

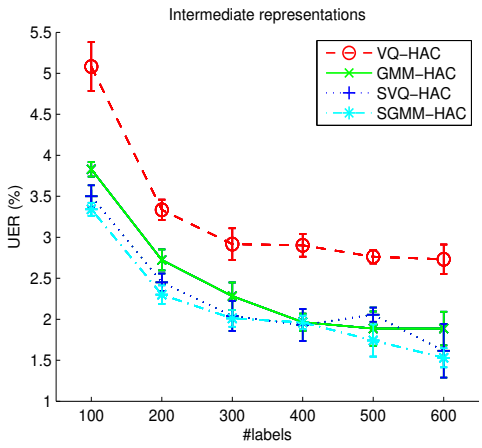

Figure 3: *Unorderered error rates (UER) obtained with different intermediate representations, using Oracle-MIDA spectral features. The horizontal axis (#labels) represents either the number of clusters N_c (for (S)VQ-HAC) or the number of GMMs N_m (for (S)GMM-HAC). The vertical bars around the data points denote the standard deviation over* 5 *different random initialisations of the NMF.*

5.1. NMF

NMF word learning was carried out using 100 multiplicative updates, and NMF-based evaluation was done using 30 multiplicative updates as in [9]. The matrix $\boldsymbol{W}$ contains 75 columns; $K = 50$ columns representing the keywords and the (non-critical) number of columns describing filler words was $(R - K) = 25$.

6. Results and Discussion

6.1. Spectral Representations

To evaluate the spectral representations, an intermediate data representation must be selected. We opt here for VQ-HAC, which was described in section 4.1, because of its limited computational demands. The assumption is thereby made that no dependency exists between the low-level features and this intermediate representation. The codebook size N_c in the creation of these intermediate representations was varied between 100 and 600, with increments of 100. The UERs that result from this experiment are shown in figure 2. These values are subject to slight variations, due to the random initialisation of the NMF framework. Therefore, each result shown in this figure is the average taken over 5 repetitions of the same experiment.

These results show that MIDA features significantly outperform the MFCCs in this word learning task. For Oracle-MIDA features, this is not unexpected since their creation involves prior linguistic knowledge. However, we can observe that with VQ-MIDA features, which are created in a completely data-driven way, we can achieve results that are competitive, especially for small codebook sizes in the VQ-HAC. This demonstrates that the effectiveness of modern discriminative spectral representations can be retained even without the use of prior knowledge.

6.2. Intermediate Representations

To evaluate the different intermediate representations described in section 4, we perform the same keyword learning experiment as described above, but now using a single spectral representation. For the spectral representation we use Oracle-MIDA features since they have proven to be the most effective in terms of recognition accuracy. The codebook size N_c and the number of GMMs N_m, for the creation of respectively (S)VQ-HAC and (S)GMM-HAC, was varied between 100 and 600, with increments of 100.

The results are shown in figure 3. We can observe in this figure that the 'soft' representations, SVQ-HAC and SGMM-HAC, outperform their hard counterparts, VQ-HAC and GMM-HAC respectively, by a fair margin. Moreover, the results show that with soft representations, the highest accuracies are reached using 600 codewords/GMMs, suggesting further improvements are still possible. This is not the case for the hard representations, as the decrease in UER seems to level off after increasing the number of GMMs/codewords beyond 500.

The difference between VQ-HAC and GMM-HAC is very large. The reason for this is that VQ-HAC makes use of Euclidean distances, assuming the clusters to be spherical, whereas the labelling in GMM-HAC is based on clusters of any shape. The differences between SVQ-HAC and SGMM-HAC, however, are not very large.

Both methods perform a soft assignment to 3 different clusters for each speech frame. For SVQ-HAC, although the the clusters are modelled by a full-covariance Gaussian, they still can only model *elliptical* shapes. The fact that SGMM-HAC may model clusters of any conceivable shape in this case turns out to be only a minor advantage performance wise. Still, its reliance on diagonal-covariance Gaussians makes it computationally much more tractable than SVQ-HAC, which requires the expensive evaluation of full-covariance Gaussians.

7. Conclusions and Future Work

The experiments of section 6.1 have shown that the use of MIDA features, as proposed in [12], leads to substantially lower UERs than the use of MFCC features. The reason is that MFCCs are only based on the heuristic idea that smooth components of the log-Mel spectrum are the most relevant for ASR, while MIDA finds a representation that optimises the discrimination between speech classes.

Moreover, we have demonstrated that MIDA representations, which are usually derived using expert knowledge, can also be determined using a data-driven clustering into speech classes with little loss of performance. This means that the goal of building a self-learning system, which precludes the use of prior knowledge, does not rule out the usage of sophisticated discriminative features used in modern ASR systems.

The experiments of section 6.2 revealed that the Vector Quantisation approach first proposed in [10], can be improved upon in several ways. Firstly, using co-occurrences of posterior features ('soft' assignments) rather than co-occurrences of single speech class labels. Secondly, using intermediate representations based on GMMs, because these allow us to model the speech clusters more accurately.

In conclusion, we have reduced the error rate obtained on this keyword acquisition task from 4.48%, obtained with VQ-HAC on MFCC features, to 1.53% with SGMM-HAC on Oracle-MIDA features. To the best of our knowledge, the latter is the best result as of yet obtained on this task. Future work will focus on a comparison with the performance obtained with a conventional ASR system, as well as developing methods to recover word-over, for example by employing sliding windows or non-negative matrix deconvolution.

8. Acknowledgements

This research is funded by KULeuven grant OT/09/028 (VASI) and the IWT-SBO project ALADIN contract 100049.

9. References

[1] J. van de Loo, J. F. Gemmeke, G. De Pauw, J. Driesen, H. Van hamme, and W. Daelemans, "Towards a self-learning assistive vocal interface: Vocabulary and grammar learning," in *Proc. SMIAE 2012*, Jeju Island, Republic of Korea, 2012.

[2] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui, "Socially embedded learning of the office-conversant mobile robot <e jijo-2 e>," in *Proc. IJCAI*, 1997, pp. 880–887.

[3] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein, "Mobile robot programming using natural language," *Robotics and Autonomous Systems*, vol. 38, pp. 171–181, 2002.

[4] M.-h. Siu, H. Gish, S. Lowe, and A. Chan, "Unsupervised audio patterns discovery using hmm-based self-organized units," in *Proc. Interspeech 2011*, Makuhari, Japan, 2011, pp. 2333–2336.

[5] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proc. Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, ser. HLT-Short '08, Stroudsburg, PA, USA, 2008, pp. 165–168.

[6] M. Huijbregts, M. McLaren, and D. A. van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Proc. ICASSP 2011*, Prague, Czech Republic, 2011, pp. 4436–4439.

[7] O. Räsänen, "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events," *Cognition*, vol. 120, pp. 149–176, 2011.

[8] K. Gold and B. S. Doniec, C. Crick, "Robotic vocabulary building using extension inference and implicit contrast," *Artificial Intelligence*, vol. 173, pp. 145–146, 2009.

[9] J. Driesen, "Discovering words in speech using matrix factorization," Ph.D. dissertation, K.U.Leuven, ESAT, Jul. 2012.

[10] H. Van hamme, "Hac-models: a novel approach to continuous speech recognition," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 255–258.

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *In IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[12] K. Demuynck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, ESAT, February 2001.

[13] M. Sun and H. Van hamme, "Unsupervised Vocabulary Discovery Using Non-Negative Matrix Factorization With Graph Regularization," in *Proc. ICASSP*, 2011, pp. 5152–5155.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*. Breckenridge, Denver, USA: MIT Press, 2000, pp. 556–562.

[15] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[16] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *In International Symposium on Music Information Retrieval*, 2000.

[17] E. Batlle, C. Nadeu, and J. A. Fonollosa, "Feature decorrelation methods in speech recognition. a comparative study," in *Proc. ICSLP*, Sidney, Australia, 1998.

[18] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[19] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 2002.

[20] R. Gray, "Vector quantization," *ASSP Magazine, IEEE*, vol. 1, no. 2, pp. 4–29, april 1984.

[21] M. Sun and H. Van hamme, "Coding methods for the nmf approach to speech recognition and vocabulary acquisition," in *IMCIC 2011*, Florida, USA, 2011.

[22] V. Estivill-Castro and J. Yang, "A fast and robust general purpose clustering algorithm," in *In Pacific Rim International Conference on Artificial Intelligence*. Springer, 2000, pp. 208–218.

[23] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. II*. New York, NY, USA: John Wiley, 1966.

[24] H. Sorenson and D. Alspach, "Recursive bayesian estimation using gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[25] "Acquisition of communication and recognition skills," http://www.acorns-project.org/, 2006-2009.

[26] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, and H. van den Heuvel, "A speech corpus for modelling early language acquisition: Caregiver," in *Proc. LREC 2010*, Malta, 2010.