# FAST WORD ACQUISITION IN AN NMF-BASED LEARNING FRAMEWORK

*Joris Driesen, Hugo Van hamme*

Department Electric Engineering-ESAT, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven
{joris.driesen,hugo.vanhamme}@esat.kuleuven.be

## Abstract

A speech recognition system that automatically learns word models for a small vocabulary from examples of its usage, without using prior linguistic information, can be of great use in cognitive robotics, human-machine interfaces, and assistive devices. In the latter case, the user's speech capabilities may also be affected. In this paper, we consider a NMF-based learning framework capable of doing this, and experimentally show that its learning rate crucially depends on how the speech data is represented. Higher-level units of speech, which hide some of the complex variability of the acoustics, are found to yield faster learning rates.

**Index Terms**: Acoustic Sub-Word Generation, Unsupervised Learning, Vocabulary Acquisition, Machine Learning

## 1. Introduction

One of the main practical applications of speech recognition technology is in spoken human-machine interaction. For the average consumer this is especially relevant for devices with complex functionality or a very small form factor, or in hands-busy situations. For people with motor impairment (especially of the upper limbs), on the other hand, such technology is also useful in voice-controlled assistive devices, even if their limited complexity requires only a small vocabulary. The difficulty with this is that motor impairments often co-occur with voice pathologies, causing a great deal of speaker variation, to the extent that speaker-independent models become unusable, e.g. [1, 2]. A system designed to recognize commands and act upon them must therefore learn and adapt to the specific speech patterns of any individual user. Standard automatic speech recognizers rely heavily on pre-programmed knowledge, making them less suitable in this context. There is need of a recognition system that can learn voice commands automatically, without using any prior linguistic knowledge and with only a minimal amount of supervision. The setting here differs from training of traditional speaker dependent hidden Markov models in that the user interface is required to learn from examples of its usage, i.e. it should acquire a vocabulary with associated meaning (actions). For example, in home automation where a user controls light, radio or TV with buttons on a remote control, the aim is to learn voice commands (as chosen by the user) and associate them with their related button presses. The setting is reminiscent of *socially guided machine learning* [3]. It also relates to how language learning infants create a mapping between acoustic patterns and objects in their surrounding environment, an operation called 'grounding', e.g. [4, 5].

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

In this paper, we make use of the vocabulary acquisition framework proposed in [6]. This framework is centered around Non-Negative Matrix Factorization (NMF) [7], and is able to learn a relatively small vocabulary of keywords from a set of speech data. The strength of this technique is that it can learn words embedded in utterances without the need of a segmentation, requiring no more supervision than the identity of the embedded keyword. We elaborate further on this below, in section 3. Making use of this framework, we focus on its capabilities to handle previously unseen words. More precisely, we want to know after how many training instances these new words can reliably be detected in a set of unseen test utterances. Intuitively, one can see that this largely depends on the representation of speech on which the learning framework operates. To investigate this, we make an analysis of the framework's performance using two different kinds of input. On the one hand, the method based on vector quantization of short speech segments, which was proposed in [6]. On the other hand, Unsupervised Acoustic Subword Units (ASWU's), derived in a way that is loosely based on the work in [8] and [9]. We explore the influence of these input representations' complexity, i.e. the number of different units they contain, on the learning rate of the NMF framework. Moreover, in the case of ASWU's, we investigate the influence of the temporal granularity of these units, i.e. their minimum duration.

This paper is organized as follows: in section 2.1, we discuss the unsupervized derivation of low- and high-level representations of speech. In section 3 we briefly review the NMF-based framework used for vocabulary acquisition. Finally, in section 4, we determine experimentally how quickly the framework learns a previously unseen keyword. We finish with some concluding remarks in section 5.

## 2. Representations of Speech Data

The input of our experiments consists of 13089 short English utterances (3 to 4 seconds) from a database recorded with the specific purpose of benchmarking keyword learning algorithms [10]. 9821 of these utterances are randomly selected to make up the train set, the 3268 remaining ones make up the test set. These utterances are spoken by 10 different speakers. Each utterance contains up to 4 keywords chosen from a total vocabulary of 50. In the front-end, 25ms frames with a frame shift of 10ms are considered, in which 22 MEL-scaled filterbank coefficients are computed. Their first and second order differences are added, yielding 66 coefficients per frame. This number is then reduced to 36 by MIDA, a mutual information based variant of LDA [11]. To enable vocabulary learning with NMF on this data, each utterance is converted to a single non-negative descriptive vector, approximately containing a weighted addi-

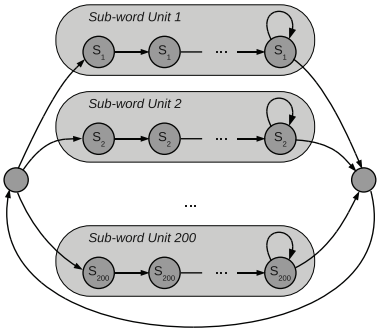


Figure 1: *The HMM lay-out which we use to train the 200 Gaussian mixture models.*

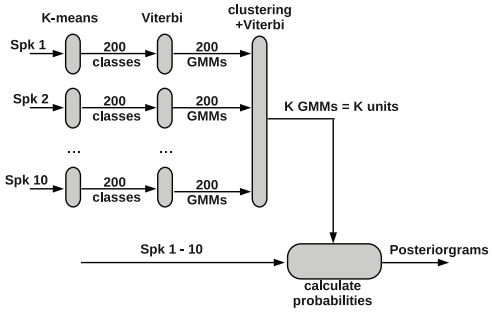


Figure 2: *An overview of the steps to obtain posteriorgrams: speaker dependent clustering in the first step, then training of GMM's based on the resulting clusters, followed by grouping the resulting GMM's across speakers and retraining them. Finally the resulting models are applied to generate posteriorgrams.*

tion of vectors that describe words. A representation that fits these conditions is a histogram of acoustic events, accumulated over the utterance's duration. Differences between representations lie in the way these acoustic events are modelled.

### 2.1. Acoustic Sub-Word Units

One of the basic precepts in this work is that prior linguistic information must not be used. Phones, along with their associated acoustic models are therefore excluded as a basis for NMF learning. In this method, we wish to define units that are conceptually similar to phones, but are derived in a data-driven way, without supervision. To this end, we first define an ergodically connected HMM as shown in figure 1. Each sub-word unit in this HMM is modelled as a sequence of states, sharing a single emission distribution. Only the last of these states is allowed to loop back to itself. As such, this topology imposes a lower limit on the duration of units in Viterbi alignment of observed utterances.

To model the emission distributions of the units, we have opted in this paper for Gaussian mixture models (GMMs), since this is the way continuous density distributions are usually modelled in ASR applications. At initialization time, these GMMs are defined with only a single mixture component, namely the Gaussians fitted over clusters of input data, determined by the Kmeans algorithm. These Gaussians are then iteratively split and re-estimated on the training data, along with their mixture weights, by means of Viterbi training. In the experiments below, state transition probabilities are not trained. They are set and kept at constant values. Most noteworthy of these constants are the loop probability on the last state of each unit, and the cross-unit transition penalties. Together these parameters play a part in the average time span of the units in a segmentation. In the experiments below, the loop probability is set to 0.95.

To ensure the discovered units are general enough to cover speech from all 10 speakers in the database equally, we follow the scheme that is shown in figure 2. Using data from each of the 10 speakers, 200 speaker-dependent acoustic sub-word units are initialized and trained, leading to a total of 2000 units, each with its own mixture model. Hierarchical clustering of these models allows us to tie them together across speakers into a total of $K$ *speaker independent* models, i.e. sub-word units. In this paper, we create two distinct sets of such units, one containing 500 of them, the other 74. The latter value is found to be the lowest possible number of models general enough to represent acoustic units, without modelling speaker characteristics.

For all the utterances of train and test set, the acoustic models and HMM topologies are used to create a directed acyclic graph, in which each arc represents a sub-word unit. From the likelihoods of the arcs in these lattices, we calculate $p(u_k|\mathbf{o_t})$, i.e. the posterior probabilities of of the units $u_k$, $k = 1...K$, given the frames of the observed input $\mathbf{o_t}$, $t = 1...T$. The resulting matrix with dimensionality $K \times T$ is called the *posteriorgram*. The smoothing effect of imposing a minimum number of states per unit can be keenly observed in this representation, as is shown in figure 3.

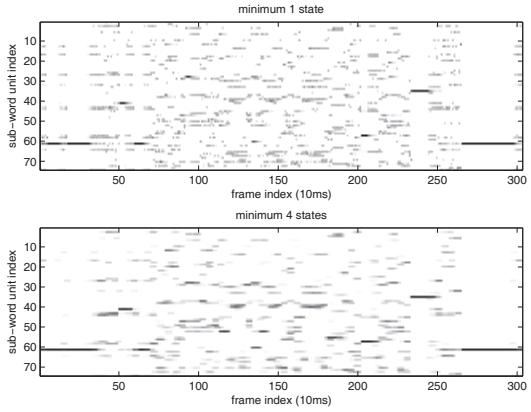


Figure 3: *Posteriorgrams of an utterance. The acoustic models here define 74 different units. The minimum length of these units is respectively 1 state (above) and 4 states (below).*

From these posteriorgrams, joint probabilities of the subword units at a time offset $\tau$ are calculated by multiplying their posteriors. We make an inherent assumption by doing this, namely

$$p(l_t = u_i, l_{t+\tau} = u_j) = p(l_t = u_i) \cdot p(l_{t+\tau} = u_j)$$

where $l_t$ denotes the unit label assigned to frame $t$. For each utterance, these joint probabilities are accumulated into a vector of length $K^2$, where $K$ is the number of subword units in the model, either 74 or 500. Vectors derived in this way contain many values close to zero, but are not sparse, making them

computationally less tractable for processing with NMF, see below. In order to introduce sparsity, all but the 3 highest posterior probabilities in each frame are set to 0, and the remaining ones are normalized to 1. Determining joint probabilities in this way is better than merely accumulating posteriors, since some of the utterance's temporal information is preserved in this operation. This idea can be taken even further by considering multiple values of $\tau$, thereby however multiplying the dimensionality of the resulting feature vectors. In our experiments, we consider time offsets $\tau$ equal to 20ms, 50ms and 90ms, yielding for every vector in train and test set a sparse vector of total length $3K^2$.

### 2.2. Vector Quantization Labels

In this method, an equal number of speech frames for every speaker is selected from the train set, yielding in total 53550 data vectors. The K-means clustering algorithm is performed to separate this data into $K$ clusters (for comparison's sake again either 74 or 500), the centroids of which comprise a vector quantization codebook. Each frame of the utterances in both train and test set can then be described by a single VQ-label between 1 and $K$. Based on a hard clustering of very short segments, without considering temporal context, they capture many fine-grained variations in the speech signal, rendering the discovery of word-related patterns more difficult.

In order for such label sequences to be used in NMF, we will then convert them into histograms. For the same reasons as mentioned above, in section 2.1, we do not accumulate label occurrences in these histograms, but the occurrences of *label combinations* at a certain time offset $\tau$. This yields sparse vectors of length $K^2$. This is in fact the same operation as the accumulation of joint probabilities from the previous section, if the label sequence is considered as a posteriorgram with a single non-zero value per column. In our experiments, we take once more the values of $\tau$: 20ms, 50ms and 90ms, leading to representations of length $3K^2$.

## 3. Vocabulary Acquisition

### 3.1. Non-Negative Matrix Factorization

For the vocabulary acquisition we make use of NMF, a paradigm in which a typically large non-negative matrix $V$ of dimensionality $M \times N$ is approximated as the rank-reducing product of non-negative matrices $W$ and $H$ which are of respective sizes $M \times R$ and $R \times N$. This factorization is solved by initializing $W$ and $H$ randomly and minimizing the cost function:

$$D(V||WH) = \sum_{ij} V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \quad (1)$$

This can be done with iterative multiplicative updates (see [7]).

Since $R$ is small compared to both $M$ and $N$, the columns of $W$ will contain patterns that approximately can be recombined by weighted additions to form the $N$ columns of $V$, with the columns of $H$ containing the appropriate weights to accomplish this. In our learning framework, where every column of $V$ is a vector that represents a speech utterance as described in previous section, the aim is for the columns of $W$ to be representations of word-like speech patterns. There are 50 keywords in the data, so $R$ should be at least equal to 50. In our experiments, $R$ is set at 75. The extra columns model acoustics that are unrelated to any of the keywords, e.g. the carrier words, noise, etc. These columns will henceforth be referred to as '*garbage columns*'. In order to learn the keywords from the train set and to detect them in the test set, we apply NMF as proposed in [6]. $V_{train}$ is augmented with a grounding part $G$ which indicates the presence of keywords in each utterance as follows:

$$G_{ij} = \begin{cases} 1 & \text{if utterance } j \text{ contains keyword } i \\ 0 & \text{otherwise} \end{cases}$$

NMF training then becomes

$$\begin{bmatrix} G \\ V_{train} \end{bmatrix} \approx \begin{bmatrix} W_g \\ W \end{bmatrix} \cdot H_{train} \quad (2)$$

Assigning keyword $i$ to a column of $W$ is then done by making the appropriate value in the corresponding column of $W_g$ large compared to all the others. As such, the 50 keyword-related columns in $W_g$ will resemble an identity matrix (or a permutation thereof) with very small off-diagonal values. The garbage columns in $W_g$ contain only very small values. Since $V_{train}$ contains data from all speakers, the models in the columns of $W$ are speaker independent.

Detection of keywords on the test set (for which the presence of keywords is unknown) is done by finding $H_{test}$ based on the acoustics of the test set and the trained $W$-matrix:

$$V_{test} \approx W \cdot H_{test} \quad (3)$$

Multiplying $H_{test}$ with $W_g$ yields an activation matrix $A$

$$A = W_g \cdot H_{test} \quad (4)$$

which is an estimate of the unknown grounding matrix of the test set, $G_{test}$. The presence of the $P_j$ keywords in utterance $j$ of the test set is predicted by identifying the $P_j$ largest elements in the corresponding column of $A$. The *unordered error rate (UER)* that results is defined as

$$UER = 100 \frac{\sum_j S_j}{\sum_j P_j} \quad (5)$$

where $S_j$ is the number of substitutions made in this prediction. The multiplicative updates in [7], used for solving NMF, only converge towards a *local* optimum, not a global one. Due to the random initialization of $W$ and $H$, the resulting factorization of NMF is therefore not deterministic. For this reason, these experiments are typically run several times, and the results are averaged over the different attempts.

## 4. Word Learning Rate

In the experiment of previous section all keywords are learned at the same time. The goal of this paper, however, is to assess the ability of this learning framework to acquire new, previously unseen words, when a vocabulary of old words is already established. To this end, the following experiment is set up. 10 of the 50 keywords in the database are selected to act as new words, after which the train data is divided into two subsets: on the one hand $V_{train}^{(1)}$, consisting of the utterances which contain any of these new keywords (4495 utterances in total), on the other hand $V_{train}^{(2)}$ containing the remaining utterances. From the latter, the 40 'old' keywords are first learned as described in section 3. Like before, all columns of $W$ are still assigned a single one of these keywords, and the number of garbage columns is still set at 25. Hence, $R$ in this training run is equal to 65.

In a next step, $W$ and $W_g$ are expanded to accommodate the 10 new keywords, and the NMF training is continued on $V_{train}^{(1)}$,

i.e. the data that contains these new words. Evaluation is done using equations 3 and 4 on all the utterances of the test set. Since this experiment mainly focuses on the 10 new keywords, we will only consider substitutions of these keywords as errors in the evaluation of the learning algorithm.

To determine how quickly new word representations are acquired, learning of the new keywords is performed on increasing random subsets of $V_{train}^{(1)}$. The larger the number of training examples, the more accurate the predictions are expected to be. As stated above, each experimental outcome is subject to a degree of randomness, which is why they are repeated four times, and their results averaged. These results are shown in figure 4.

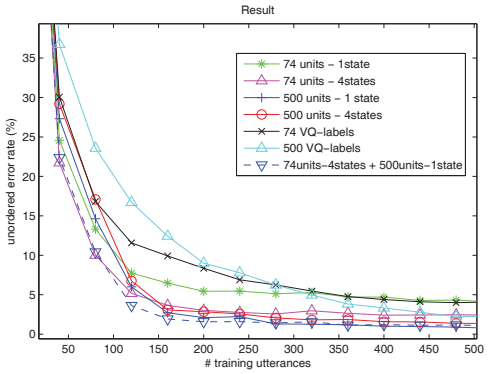


Figure 4: *The learning rate for the unseen words for increasing train sets, expressed in number of errors using the optimal detection threshold.*

We can see that the number of distinct units, as well as the way they are modelled, have a profound effect on both the learning rate and the accuracy at which the learning saturates. A higher number of units always leads to better results, regardless of what these units are. This is expected, since more units allow for more accurate word models. Sub-word units of 1 state are conceptually very similar to soft VQ, i.e. a representation in which each frame is not described a single cluster label, but with probabilistic weights for all clusters [12]. When the number of units is 74, this leads to very similar error rates, although VQ-labels are a bit quicker to learn, likely because the few realizations of each word within a small set of training utterances there show less variation. The reverse is true for 500 unit models. Since the 500 corresponding data clusters are smaller, slight acoustic differences lead more easily to different labels. Variation in the speech signal is thus much more reflected in the VQ-labels, causing learning to go exceptionally slow. This is to a much lesser extent the case for GMM's, because their complexity allows them to model such slight variations into the same sub-word unit. This way, small variations are hidden from the NMF learning framework. The same effect is perceived in comparing the 74 unit models constrained to a duration of a single frame, with those constrained to 4 frames. Fine-grained variations are here too prevented from showing up in the label sequence because of this longer minimum duration. Imposing such a longer minimum duration in a model of 500 units harms performance, when compared to unconstrained units, likely because the former is able to model stationary parts of the signal that are shorter than 40ms, whereas the latter is not.

It has been shown in the past that NMF can very conveniently combine knowledge sources [13]. With this idea in mind, an experiment was set up in which we combine the data representation enabling the fastest learning (74 units - 4 states) with the one enabling the lowest UER. The result, shown as a dashed curve in figure 4, shows that this hybrid data representation has the advantages of both, facilitating fast learning as well as high accuracies.

## 5. Conclusion

In this paper, we have discussed several means of representing a speech utterance as input data to a NMF-based learning framework for vocabulary acquisition. These input types include very low-level representations of the acoustics, as well as higher-level representations called acoustic subword units, similar to phones. We investigated the influence of the type of input on the speed with which new words are acquired by the learning framework. During the learning process, no use was made of prior linguistic knowledge and supervision was kept to a minimum. This enables e.g. the learning of voice commands from speech that is strongly affected by various kinds of voice impairments. The number of training examples necessary for word learning is of great importance in such applications. We have shown experimentally that learning is most quickly performed with higher-level data representations, and that doesn't necessarily come at the cost of reduced performance. Future work includes evaluation of such representations on larger vocabularies.

## 6. References


[1] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices," *IEEE Engineering in Medicine and Biology*, vol. 16, no. 4, pp. 74–82, 1997.

[2] O. Saz, C. Vaquero, E. Lleida, J. M. Marcos, and C. César, "Study of maximum a posteriori speaker adaptation for automatic speech recognition of pathological speech," in *Proc. Jornadas en Tecnología del Habla*, 2006.

[3] A. L. Thomaz and C. Breazeal, "Transparency and socially guided machine learning," in *Proc. ICDL 2006*, Bloomington, Indiana, USA, 2006.

[4] D. Roy, "Learning visually-grounded words and syntax for a scene description task," *Computer Speech and Language*, 2002.

[5] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[6] H. Van hamme, "Hac-models: a novel approach to continuous speech recognition," in *Interspeech 2008*, Brisbane, Australia, 2008.

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

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

[9] M.-H. Siu, H. Gish, A. Chan, and W. Belfield, "Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision," in *Proc. Interspeech 2010*, Makuhari, Chiba, Japan, 2010.

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

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

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

[13] H. Van hamme, "Integration of asynchronous knowledge sources in a novel speech recognition framework," in *In Proc. ITRW on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, 2008.