COMBINING NDHMM AND PHONETIC FEATURE DETECTION FOR SPEECH RECOGNITION

Torbjørn Svendsen and Jarle Bauck Hamar[*]

Department of Electronics and Telecommunications
NTNU
Trondheim, Norway

torbjorn@iet.ntnu.no, jarle.hamar@gmail.com

ABSTRACT

Non-negative HMM (N-HMM) [1] is a model that is well suited for modeling a mixture of e.g. audio signals, but does not have the ability to generalize to model unseen data. Non-negative durational HMM (NdHMM) has recently been proposed [2] as a modification to N-HMM that can allow for generalization, and thus make the approach suitable for automatic speech recognition. A detector-based approach to speech recognition has been studied by several researchers as an alternative to the traditional HMM approach. A bank of phonetic feature detectors will produce phonetic feature posteriors, which fit well with the non-negativity constraint of NdHMM. We review the NdHMM approach proposed in [2] and propose to extend this approach by combining NdHMM with a phonetic feature detection front-end in a tandem-like system. Experimental results of the proposed approach are presented.

Index Terms— ASR; Non-negative durational HMM; Phone recognition; Phonetic feature detection

1. INTRODUCTION

Non-negative matrix factorisation (NMF) has been shown to be useful in various disciplines. A key propery of NMF is the ability to extract latent components from data, yielding a reduced rank approximation of the non-negative matrix as an additive combination of the latent components.

Although this decomposition is inexact, the reduced rank approximation has been shown to be useful in many applications. In [3], NMF has been successfully used to discover phone patterns by representing each utterance in the database using weighted phone lattice transition probabilities in the columns of $\mathbf{V}$. Further, in [4], convolutional NMF (cNMF) [5] was used to discover phone structures.

Probabilistic extensions for NMF allow the use of sophisticated statistical techniques while still using the general ideas of NMF. In [6], a probabilistic extension of NMF is presented for modeling sound spectrograms. The columns of a spectrogram $\mathbf{V}$ are modeled as histograms of "sound quanta". The amount of sound quanta in a given time-frequency bin is indicated by the Fourier magnitude of that bin. After a normalization, the spectrogram can be considered a joint probability distribution $P_t(f)$ over time and frequency and is represented as follows:

$$P_t(f) = \sum_z P(f|z)P_t(z). \tag{1}$$

This formulation states that a quantized version of the spectrogram can be generated by performing multiple draws from the distribution $P_t(f)$. Each draw adds a sound energy quantum to the corresponding time-frequency bin. The distribution, $P_t(f)$, is defined as a linear combination of a set of time independent dictionary components ($P(f|z)$) weighted with time dependent weights ($P_t(z)$) and are represented using multinomial distributions. Thus, each time frame of the spectrogram is generated by performing multiple draws: first a dictionary component, z, is selected according to $P_t(z)$, then the frequency to be assigned an additional energy quantum is chosen according to $P(f|z)$. The draws continue until the frame energy matches the observed energy. The mixture weights of the model therefore capture the temporal variation in the input signal. The formulation in Equation (1) is also referred to as probabilistic latent semantic analysis (pLSA) in the literature [7].

A major limitation in using the above formulation for modeling speech is that the spectrogram is represented by a single set of dictionary components, $P(f|z)$. This limits the expressive power of the model as the speech spectrum is non-stationary. In [8], [1] and [9] it has been shown that HMMs can be combined with NMF to incorporate the non-stationary component as a Markov model which allows changing the dictionary components with time. This model is called non-negative HMM (N-HMM)

In the N-HMM, each state q has a fixed set of dictionary components $P(f|z,q)$ with time varying weights $P_t(z|q)$. Thus N-HMM is able to describe different parts of the input signal with different states. For a speech signal the states may correspond to different phones. In [8], it has been shown that the model can extract phone structures.

Although N-HMM has been reported to be successful for separating mixture signals, such as music or speech and noise, it is not suited for modeling components of a speech recognition system (ASR). This is because the weights $P_t(z|q)$ are dependent on the absolute time t in the utterance, making the model unable to generalize to unseen data. In this paper, we summarize a modified approach to the N-HMM formulation, N-d HMM, for use in ASR set proposed in [2], [10] where the idea is to use the same set of weights with every visit to the state and force the variation on the weights to be dependent on the duration for which the state is active instead of absolute time.

In [2], the processing was based on a spectrogram or a modfied MFCC representation as the input. An alternative to tradtitional approaches is to base the recognition process on a bottom-up detection [11], [12]). In this approach, we use a bank of detectors to find estimates of the posterior probabilities of a set of phonetic features. We propose to use the resulting posteriorgram as input to an NdHMM

[*] Work performed while a PhD fellow at NTNU

Algorithm 1 Generative process of NdHMM

```
Draw $q_1$ from $P(q_1)$
Set $d_1 = 0$
for $t = 1 \rightarrow T$ do
    Draw $v_t$ from $P(v_t|q_t)$
    Draw $z_t$ from $P(z_t|q_t, d_t)$
    for $v = 1 \rightarrow v_t$ do
        Draw $f$ from $P(f|q_t, z_t)$
        $\mathbf{x}_t[f] = \mathbf{x}_t[f] + 1$
    end for
    Draw $q_{t+1}$ from $P(q_{t+1}|q_t)$
    if $q_{t+1} = q_t$ then
        $d_{t+1} = \min(d_t + 1, D_{q_t})$
    else
        $d_{t+1} = 0$
    end if
end for
```

recognizer in a tandem system [13] approach. We present details of the approach, and experimental results using both the system in [2] and the proposed tandem N-d HMM system for phone recognition.

2. NDHMM

The N-HMM combines NMF and HMM to allow the speech signal to be expressed with time varying dictionary components. The N-HMM described in [8] is a hidden Markov model where the distributions governing the observations generated by a state q are given by $P_t(f|q) = \sum_z P(f|q,z)P_t(z|q)$, i.e. a set of dictionary components $P(f|q,z)$ and a set of time dependent weights, $P_t(z|q)$. The observation vectors are generated from the model by performing multiple draws from the underlying distribution and for each draw adding an energy quantum to the corresponding time-frequency bin. The number of draws, v_t, for each state at time t is explicitly modeled using a Gaussian distribution $P(v_t|q_t)$, the energy distribution of the state.

It is important to note that the weights $P_t(z|q)$ are time dependent. The variation of weights for each time frame captures the temporal variations in the input signal although the dictionary components are always time invariant, only conditioned on the state. However, this implies that the weights are *utterance dependent*, and the model is not particularly useful for modeling unseen data. We cannot remove the time dependency of the weights as they capture the temporal dynamics. Further, having constant weights for all time frames, will collapse the multinomial mixture models to a single multinomial for each state and will result in a poor model.

In order to overcome the above problem, we propose to make the weights dependent on the state occupancy duration, rather than absolute time. Thus, the same set of weights are used every time the state is visited. By denoting the current duration of the current state q_t as d_t, the weight distribution for the state is time independent, but conditioned on d_t, i.e. $P_t(z|q) \rightarrow P(z_t|q_t, d_t)$. This modification does not impose any limit to the state occupancy duration, i.e. the process may be in a specific state indefinitely long, consistent with the Markov property of HMM.

Estimating $P(z_t|q_t, d_t)$ for an infinite number of durations is impossible. To alleviate this problem, we introduce a threshold D_q for the duration counter d. If the process has been in state q for $t \geq D_q$, we assume that the weight distribution does not change, i.e. $P(z_t|q_t, d_t) = P(z_t|q_t, D_q)$ for $t \geq D_q$. The proposed modification removes the time dependency problem of the weights, making the modified structure, which we refer to as Non-negative durational HMM (NdHMM), suitable for speech recognition tasks.

In an effort to increase the discriminative capability of the model, all the draws were restricted to be from one dictionary component; i.e only one dictionary component is drawn for each time frame, and the entire output vector is created using several draws from that component. The generative process of the NdHMM for creating the output sequence $\bar{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ is given in Algorithm 1.

2.1. Parameter Estimation

The complete set of parameters in the NdHMM are:

- The dictionaries: multinomial distributions: $P(f|z,q)$
- The weights: multinomial distributions $P(z|q,d)$
- Energy distributions: Gaussian distributions $P(v|q)$
- Transition probabilities: Markov model $P(q_{t+1}|q_t)$
- Initial state probabilities: $P(q_1)$

As in standard HMM, the input vectors are assumed to be conditionally independent, and the transition probabilities are only dependent on the current state.

To estimate the parameters of the NdHMM, we use a similar approach to what is described in detail in [8]. The training procedure uses the EM-algorithm for updating the parameters of the NdHMM, similar to training of the conventional HMM.

The complete data log likelihood becomes:

$$\begin{aligned}\log P(\bar{\mathbf{x}}, \bar{\mathbf{z}}, \bar{\mathbf{q}}, \bar{\mathbf{v}}) &= \log P(q_1) + \sum_{t=1}^{T-1} \log P(q_{t+1}|q_t) \\ &+ \sum_{t=1}^{T} \log P(v_t|q_t) + \sum_{t=1}^{T} \log P(z_t|q_t, d_t) \\ &+ \sum_{t=1}^{T} \log P(\mathbf{x}_t|z_t, q_t) \end{aligned} \tag{2}$$

where $\bar{\mathbf{x}}$, $\bar{\mathbf{z}}$, $\bar{\mathbf{q}}$ and $\bar{\mathbf{v}}$ denotes the sequence of feature vectors, dictionary components, states and energy respectively. In order to estimate a new set of distributions $\hat{P}(\cdot)$ based on the current estimates $P(\cdot)$, the first step of the EM-algorithm in to estimate:

$$\begin{aligned}\mathcal{L} &= E_{\bar{\mathbf{z}},\bar{\mathbf{q}}|\bar{\mathbf{x}},\bar{\mathbf{v}}}\{\log \hat{P}(\bar{\mathbf{x}}, \bar{\mathbf{z}}, \bar{\mathbf{q}}, \bar{\mathbf{v}})\} \\ &= \sum_{\bar{\mathbf{q}}} \sum_{\bar{\mathbf{z}}} P(\bar{\mathbf{z}}, \bar{\mathbf{q}}|\bar{\mathbf{x}}, \bar{\mathbf{v}}) \log \hat{P}(\bar{\mathbf{x}}, \bar{\mathbf{z}}, \bar{\mathbf{q}}, \bar{\mathbf{v}}) \end{aligned} \tag{3}$$

In the second step of the iteration, (3) is maximized with respect to the new parameters. The resulting update formulae for the parameters are given in [2], where a more detailed description of the derivation of the formulae is provided.

3. PHONETIC FEATURE DETECTION

An alternative to the standard approach to speech recognition, is to base the recognition process on a bottom-up, detection based approach (see e.g. [11], [12]). Our version of this approach is to design a front-end that consists of a bank of phonetic feature detectors. The task of each feature detector is to estimate the posterior probability of a phonetic feature being active for a given speech frame. The output of the bank of feature detectors are combined in a feature vector, and subsequently used as input to the NdHMM based phone recognizer. Since the output of the bank of feature detectors is a set of phonetic feature posteriors, using this as the feature vector will satisfy the non-negativity requirement of NdHMM. In our approach, we have used a binary phonetic feature set consisting of manner and place features, augmented by Chomsky and Halle distinctive features. The feature set has 22 phonetic features and is defined in [14].

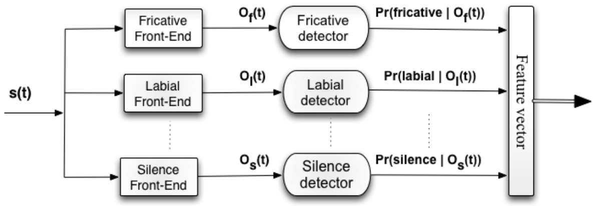

Fig. 1: *Detector bank (adapted from [12])*

3.1. Detector training

We are building a bank of detectors, comprising individual detectors for each target (phone or phonetic feature). Thus, each detector needs to be trained individually. In principle, the feature vectors presented to each detector can be tailored to maximize the performance of the individual detectors, as illustrated in figure 1. Here, we have however used the same feature extractor for all detectors.

Each branch in the detector bank is implemented as shown in figure 2, which is the same basic detector structure as used in [12]. The speech is analyzed by a 23 channel mel filterbank producing band energy estimates from 25ms windows every 10 ms. Split temporal context feature vectors [15] are then produced using a context window of 15 frames in forward and backward direction. This provides information on the temporal evolution to the detection process. The left and right context feature vectors are used for training two independent ANNs to estimate phonetic feature posteriors. The output of these two ANNs are subsequently used as input to a merger ANN, that combines the information to produce the final feature posterior.

The artificial neural networks used in the detectors are all multilayer perceptrons (MLPs) with a single hidden layer of 500 nodes.

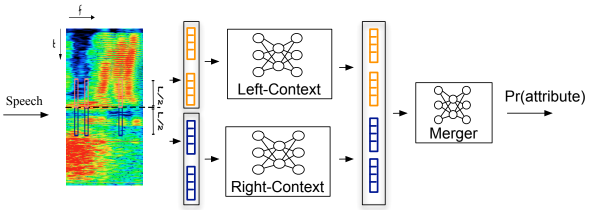

Fig. 2: *Phonetic feature detector architecture (adapted from [12])*

4. EXPERIMENTS

We experimented with the proposed NdHMM model for a phone recognition task using the TIMIT speech corpus [16]. The data is divided into a test set (1344 utterances), a training set (3296 utterances) and a development set (400 utterances).

The manual phonetic labeling is used for testing and for training of the models, after appropriate mapping from the 61 phone set to the 39 phone set [17].

Note that we did not use a language model (e.g. phone bigram) in any of the systems investigated.

NdHMM requires the features to be non-negative. If we are to use speech representations that do not satisfy this requirement, such as mel-frequency cepstral coefficients (MFCCs) it will require a transformation to make them non-negative. Using the spectrogram as features is a good alternative as the values are non-negative and do not require any transformation. The only drawback is that they have high dimensionality, requiring a high number of parameters to be estimated, and thus more computer time for training. A transformation that will produce non-negative features is the logistic sigmoid:

$$f(x;\alpha,\mu_x)=\frac{1}{1+\exp\{-\alpha(x-\mu_x)\}} \tag{4}$$

where $f(\mu_x)=0.5$ and α controls the slope. We have found that with a proper choice of the slope factor, the sigmoid transformation has only marginal influence on the recognition performance for conventional HMMs, indicating that the transformation should not have a severe adverse impact on the recognition rate in general.

NdHMM training is initialized by performing K-Means clustering on the data from each phone. The resulting cluster centroids are then used as the initial dictionary components. The initial weights are set constant with regard to the duration and their values reflect the amount of data in each cluster. The duration threshold is set individually for each state to cover at least 90 % of the durations seen in the training data.

4.1. MFCC features

13 MFCC's including C_0 were extracted using a 25 ms window with 10 ms shift and a 26 channel mel filterbank. In NdHMM, energy is modeled as the sum of the components in the input vector. The feature vector presented to NdHMM excludes C_0 and is transformed to be non-negative by the sigmoid transformation. The C_0 information is preserved by scaling the transformed feature vector to sum to C_0.

4.2. Filterbank parameters

Using filter bank energies has some inherent advantages, particularly that the coefficients are non-negative so that no transformation is necessary. Furthermore, the representation has a clear physical interpretation. The filter bank analysis in our experiments calculated log power channel estimates from a 26 channel mel filter bank using a 25ms Hamming window shifted 10ms per analysis frame.

The NdHMM formulation assumes a multinomial distribution which in turn implies that the input signal is integer. It turns out that the performance of the system is quite sensitive to the dynamic range of the input spectrogram. We applied a scaling to the input signal:

$$\mathbf{X'}_i=[\gamma_f\cdot\mathbf{X}_i/X_{max}] \tag{5}$$

where X_{max} is the largest spectrogram component in the utterance from which $\mathbf{X}_i$ is extracted and $[\cdot]$ denotes rounding and γ_f is a scaling constant chosen by experimentation (typically $\gamma_f\approx 125$).

4.3. Acoustic feature posteriorgram parameters

An alternative to the spectrogram as a naturally non-negative speech representation, is a posteriorgram. Such a representation can be constructed e.g. by using the output of the bank of phonetic feature detectors described in section 3. When the posteriorgram from the phone detectors is generated using split temporal context features, we also know that some dynamic information will be taken into consideration when the posterior estimates are calculated.

A posteriorgram generated from the TIMIT sentence fadg0_si649, "It suffers from a lack of unity of purpose and respect for heroic leadership." is shown in Figure 3, where we have zoomed in on the section "a lack of unity". The red lines in the plot show the ideal phonetic feature activations.

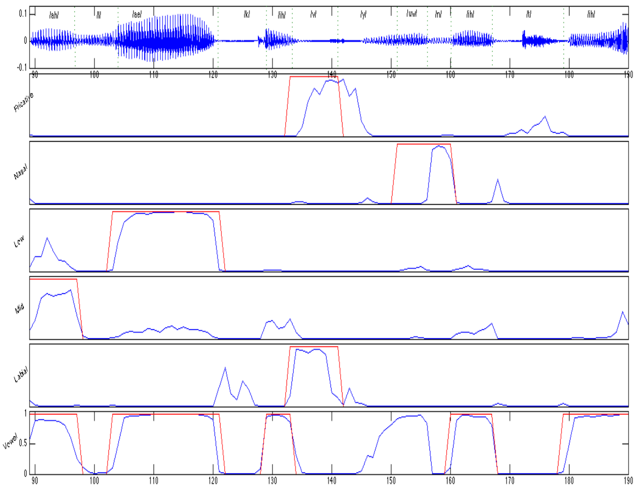

Fig. 3: *Posteriorgram of selected features from excerpt of sentence fadg0_si649, "a lack of unity". Stylized activations generated from the manual labeling are shown in red. The phonetic features are (top to bottom): Fricative, Nasal, Low, Mid, Labial, Vowel.*

As can be observed from Figure 3, the detectors will output fairly smooth posterior estimates which in many cases do not reach a steady activation level. This implies that there may be exploitable information in the temporal evolution of the posterior estimates, and that e.g. computing first and second order temporal derivatives may give performance improvements.

The feature values of a posteriorgram lie between 0 and 1. Initial experiments showed that this representation was not a good choice. Using log values would give a better dynamic range, but would of course necessitate handling negative values. After some experimentation we chose to use the following input transform:

$$\mathbf{X}' = \log \mathbf{X} - \log(X_{min})$$
$$\hat{\mathbf{X}} = [(\gamma_p \cdot (\mathbf{X}'/X'_{max})] + 1$$

where X_{min} is the smallest value of the posteriorgram of an utterance and X'_{max} is the largest value in the transformed log posteriorgram. $[\,\cdot\,]$ denotes rounding and γ_p is a scaling constant chosen through experimentation (typically $\gamma_p \approx 100$). The first step of the feature transformation takes the log of the posteriorgram, and adds a positive value to the resulting features to ensure that the result is non-negative. The second step merely scales the features to be in the range $1 \leq \hat{X} \leq (\gamma_p + 1)]$.

4.4. Results

We trained baseline HMM systems using 3-state context independent phone models with 16 and 32 component GMM emission densities. The first system used 13 MFCC coefficients including C_0. The second system added first and second order derivatives to produce a 39-dimensional feature vectors. The performance of the baseline systems is given in the first four rows of Table 1.

NdHMM systems were then trained for various types of feature vectors. All these systems had a dictionary size of 32.

The first system used 26 mel filter bank log energies as input, and achieved an accuracy of 48.5%, which is comparable to the HMM baseline with only static input features. A system based on transformed MFCC parameters as described in 4.1 was also trained, but had a performance inferior to the filterbank system.

We then trained a system using the output of the phonetic feature detectors, transformed as described in section 4.3 as feature vectors. The output of the individual detectors are the posterior estimates of the target and the anti-target classes respectively. Although the estimates do not exactly sum to one, it turns out that we can reduce the dimensionality of the feature vector from 44 (posteriors of both target and anti-target) to 22 (target posteriors only) without loss of performance. This system obtained an accuracy of 60.6% on the TIMIT test set.

This is a huge improvement over using mel filterbank log energies as input. A possible interpretation could be that the posteriorgrams include some dynamic information due to the split temporal context feature extraction which employs a 15 frame window in both forward and backward direction. We thus trained a standard HMM system using the posteriorgram as input. The system used mean and variance normalization of the log posteriors. Using only static posteriors, this system performed a little better than the baseline using 13 static MFCC coefficient feature vectors, and had a performance of 51.5 % accuracy when using 32-component GMMs for the observation densities. Adding first and second order dynamic features improved the performance further to 55.5% accuracy. This indicates that even though the posteriorgram may contain some dynamic information, a significant part of the temporal information must still be extracted by other means.

Standard HMM systems benefit greatly from including dynamic information in the feature vector. We see from Table 1 that adding dynamic features increases accuracy by around 15% absolute. Note that all the results reported in the table are obtained without a phone language model. Using a bigram phone model will typically result in a performance improvement of 3.5-4% absolute. Although some temporal information is inherent in the posteriorgram due to the split temporal context feature extraction, we saw above that there is additional temporal information in the posteriorgram that should be possible to exploit. However, we need to make the derivatives non-negative to use them with NdHMM. As an initial experiment, we computed the first and second order derivatives from the standard phonetic feature posteriorgram. The derivatives were then transformed to non-negative values using the logistic sigmoid of Eq (4), setting μ_x to the utterance mean for the respective derivatives. The resulting system exhibits a 1.9% absolute accuracy improvement at the cost of a parameter increase by a factor of 3.

The rightmost column of Table 1 shows the number of parameters per phone model for the different systems. The NdHMM systems do in general require less parameters than the HMM counterpart, without significant loss of performance.

5. DISCUSSION

The performance of the tandem system is clearly dependent on the performance of the phonetic feature detectors. Using the phonetic feature posteriors as basis for frame level MAP classification, we observed that the frame accuracies for the detectors were typically

Table 1: Phone recognition performance on TIMIT test set. The number affixed to the HMM system is the GMM size. NdHMMs have dictionary size 32. Feature type PF means phonetic feature posteriors. Context independent models, no language model.

System	Feature type	Feature dim	Accuracy (%)	#param's/ phone
HMM-16	MFCC_0	13	48.2	1296
HMM-32	MFCC_0	13	49.3	2592
HMM-16	MFCC_0_D_A	39	62.6	3792
HMM-32	MFCC_0_D_A	39	64.9	7584
HMM-32	PF	22	51.5	4320
HMM-32	PF+$\Delta + \Delta\Delta$	66	55.5	12768
NdHMM	F-bank	26	48.5	1222
NdHMM	PF	22	60.6	1034
NdHMM	PF+$\Delta + \Delta\Delta$	66	62.5	3102

around 90-95%. Note however that a chance classifier (always choosing the alternative with the highest prior) gave 80-90% accuracy for most features, and even non-informative priors would produce 50% accuracy. Regarding the frame classification as an information retrieval problem, we saw that the detectors had F-scores in the range 0.5- 0.95, i.e. a fairly wide range. Fortunately, the performance for the most frequent features were (with exceptions for "Tense", "Coronal" and "Back") in the high end. However, we are confident that improving the detector performance will be important for improving the overall system performance.

6. CONCLUSIONS

We have presented the principles of Nd-HMM and demonstrated that this technique, when combined with the bottom-up approach of creating feature vectors based on phonetic feature posterior estimates, can achieve a performance that is comparable to standard HMM on a phone recognition task. Furthermore, this performance is achieved with significantly fewer parameters than HMM, as indicated by the last column of Table 1. The approach has promise, and we believe that with better representation of the dynamic information the performance can be further improved.

7. ACKNOWLEDGEMENTS

This work has been supported by the Research Council of Norway through the projects "Spoken Information Retrieval by Knowledge Utilization in Statistical Speech Processing - SIRKUS" and "Atomic Units for Language Universal Speech Processing - AULUS". The first author wishes to thank prof. Jim Glass and the MIT CSAIL Spoken Language Systems Group for hosting him during a major part of this work.

8. REFERENCES

[1] G. Mysore, P. Smaragdis, and B. Raj, "Non-negative HMM of Audio with Application to Source Separation," in *Proc. Latent Variable Analysis and Signal Separation*, St. Malo, France, 2010, pp. 140–148.

[2] J. B. Hamar, R. S. Doddlipata, T. Svendsen, and T. V. Sreenivas, "Non-Negative Durational HMM ," *Proc. 2013 IEEE Int'l Workshop on Machine Learning for Signal Processing*, Sept. 2013.

[3] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering Phone Patterns in Spoken Utterances by Non-Negative Matrix Factorization," *IEEE SP Letters*, vol. 15, pp. 131 –134, 2008.

[4] P. D. O'Grady and B. A. Pearlmutter, "Discovering Speech Phones Using Convolutive Non-negative Matrix Factorisation with a Sparseness Constraint," *Neurocomputing*, vol. 72, pp. 88–101, Dec. 2008.

[5] P. Smaragdis, "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," in *Proc. Independent Component Analysis and Blind Signal Separation*, Granada, Spain, 2004, pp. 494–499.

[6] B. Raj and P. Smaragdis, "Latent Variable Decomposition of Spectrograms for Single Channel Speaker Separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*, oct. 2005, pp. 17 – 20.

[7] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, pp. 177–196, 2001.

[8] G. J. Mysore, *A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures*, Ph.D. thesis, Stanford University, 2010.

[9] G.J. Mysore and P. Smaragdis, "A Non-Negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics," in *ICASSP 2011*, may 2011, pp. 17 –20.

[10] Jarle Bauck Hamar, *Using Sub-Phonemic Units for HMM Based Phone Recognition*, Ph.D. thesis, NTNU, 2013.

[11] C.-H Lee and S.M Siniscalchi, "An Information-Extraction Approach to SpeechProcessing: Analysis, Detection, Verification, and Recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, May 2013.

[12] S. M. Siniscalchi, D.-C. Lyu, T Svendsen, and C.-H Lee, "Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data," *IEEE Trans on Audio. Speech and Language Processing*, vol. 20, no. 3, pp. 875–887, Mar. 2012.

[13] H Hermansky, D Ellis, and S Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1635–1638 vol.3, June 2000.

[14] S.M. Siniscalchi, "Combining Speech Attribute Detection and Penalized Logistic Regression for Phoneme Recognition," *Neurocomputing*, vol. 93, pp. 10–18, Sept. 2012.

[15] P Schwarz, P Matejka, and J Cernocky, "Hierarchical Structures of Neural Networks for Phoneme Recognition," *Proc. ICASSP*, vol. 1, pp. I–I, 2006.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27403–+, Feb. 1993.

[17] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641 –1648, nov 1989.