

# On the Relevance of Auditory-Based Gabor Features for Deep Learning in Robust Speech Recognition

Angel Mario Castro Martinez<sup>a,b,\*</sup>, Sri Harish Mallidi<sup>c</sup>, Bernd T. Meyer<sup>c</sup>

<sup>a</sup>Department für medizinische Physik und Akustik, Carl von Ossietzky Universität Oldenburg, Germany

<sup>b</sup>Exzellenzcluster Hearing4all, Germany

<sup>c</sup>Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

## Abstract

Previous studies support the idea of merging auditory-based Gabor features with deep learning architectures to achieve robust automatic speech recognition, however, the cause behind the gain of such combination is still unknown. We believe these representations provide the deep learning decoder with more discriminable cues. Our aim with this paper is to validate this hypothesis by performing experiments with three different recognition tasks (Aurora 4, CHiME 2 and CHiME 3) and assess the discriminability of the information encoded by Gabor filterbank features. Additionally, to identify the contribution of low, medium and high temporal modulation frequencies subsets of the Gabor filterbank were used as features (dubbed LTM, MTM and HTM respectively). With temporal modulation frequencies between 16 and 25 Hz, HTM consistently outperformed the remaining ones in every condition, highlighting the robustness of these representations against channel distortions, low signal-to-noise ratios and acoustically challenging *real-life* scenarios with relative improvements from 11 to 56% against a Mel-filterbank-DNN baseline. To explain the results, a measure of similarity between phoneme classes from DNN activations is proposed and linked to their acoustic properties. We find this measure to be consistent with the observed error rates and highlight specific differences on phoneme level to pinpoint the benefit of the proposed features.

## Keywords:

Auditory features, spectro-temporal processing, deep neural networks, automatic speech recognition .

## 1. Introduction

Over the last decade there have been major advances in automatic speech recognition (ASR), which mainly have promoted ubiquitous speech enhanced technologies in our daily lives.

The approaches to transcribe speech into words have changed in several ways throughout the years. Not even 10 years ago, the use of hidden Markov models (HMM) to represent speech as sequence of time-varying states and Gaussian mixture models (GMM) to statistically fit the acoustic input to these states was widely adopted as the standard for ASR; actually, due to the implementation of discriminative training methods to optimize HMM classification (Povey, 2005), (He et al., 2008), GMM-HMM recognizers yielded the best performance among other systems.

Recently, however, deep neural networks (DNNs) have successfully replaced GMMs in both small and large vocabulary tasks (Mohamed et al., 2011, 2012), (Pan et al., 2012), (Seide et al., 2011), (Sainath et al., 2011) for reasons better explained in (Hinton et al., 2012).

In spite of all aforementioned advances, ASR performance still lags far behind its human counterpart, especially in noisy and reverberant environments, thereby preventing the further development of technologies empowered by ASR, regardless how appealing or necessary they might be. To bridge this gap, researchers have focused on two different, but not mutually exclusive, strategies: developing better back-ends and extracting more informative discriminable features. For a thorough overview of noise-robust techniques successfully implemented in ASR research, refer to (Li et al., 2014).

Concerning DNNs, one recipe to accomplish the former goal is relatively straightforward, it involves the trade-off between lowering the error and generalization (much like many other machine learning algorithms) and depends on how the system performs on a cross-validation set. On the one hand if the purpose is to minimize the loss function, the course of action is to increase the model complexity, i.e. increase the number of parameters either the number of neurons per layer or the depth of the network by adding additional layers.

On the other hand if the loss function of the cross-validation set increases (a situation known as overfitting), there is a need for more training examples. In some cases the amount of available data is limited and despite the vast advances in computing software and hardware, training times do not scale well on deep architectures; for those cases, the second objective seems to be more accessible. As the healthy human ear is still unmatched in

☆  This work is licensed under a “CC BY-NC-ND 4.0” license.

Email addresses: angel.castro@uni-oldenburg.de (Angel Mario Castro Martinez), mallidi@jhu.edu (Sri Harish Mallidi), bernd.t.meyer@jhu.edu (Bernd T. Meyer)

<sup>1</sup>Corresponding author

its robustness (Lippmann, 1997), (Scharenborg, 2007), (Meyer and Kollmeier, 2010), mimicking its principles improves existing feature extraction methods for ASR; better representations, in turn, could potentially lead to a broader understanding of the underlying principles of human auditory processing.

The use of feature extraction techniques inspired by the auditory system have previously demonstrated a boost in speech recognition performance. Even the most widely used Mel frequency cepstral coefficients (MFCC) or features resulting from the perceptive linear predictive (PLP) analysis of speech (Hermansky, 1990), intrinsically implement biological findings. Owing to the glottal source of speech low frequencies have more energy therefore a pre-emphasis stage equalizes the signal power; both features use a different scale for frequency warping derived from psychoacoustic measurements (the Mel scale for MFCC and the Bark scale for PLP). Non-linear functions are applied for amplitude compression mimicking loudness perception of the auditory system (the logarithm for MFCC and an intensity-loudness power law for PLP features, respectively). Additionally, in PLP several properties of hearing concerning asymmetries in frequency selectivity and equal loudness are simulated in more detail resulting into a closer auditory-like spectrogram than the log-Mel used in MFCC.

In order to increase the recognizer robustness to channel distortions and other convolutional noise sources, MFCC and PLP features were extended by processing mechanisms such as cepstral mean normalization and RASTA processing (Hermansky and Morgan, 1994), the latter consists of bandpass filtering the compressed spectral amplitudes to emphasize transients, imitating the auditory periphery tendency to focus on the relative values of an acoustic input.

Conversely, temporal evolution of specific spectral energy bands has been captured by temporal patterns (TRAPS) and hidden activation TRAPS (HATS) (Hermansky and Morgan, 1994) features to detect underlying phonetic class structures usually taking long-time segments (1 second) compared to spectral analysis (10 ms). The hypothesis grounding the development of TRAPS and HATS suggest the spectral information perceived by the human auditory system serves not as classifier but as a frequency sub-band selector of the region most dominated by the target signal and thus temporal analysis of such bands is how the utterance is decoded in the cortex.

Kim and Stern (2009), proposed an algorithm to calculate power normalized cepstral coefficients (PNCC) as an alternative to the conventional MFCC. The calculation of PNCC integrates a Gammatone filterbank to better approximate the place-frequency mapping of the basilar membrane (Patterson et al., 1992) as opposed to the triangular filters form MFCC, it also replaces their logarithmic non-linearity with a power function derived from physiological observations of auditory nerve firings to fit the dynamic dependency of the input sound level and the perceived loudness used to compress the output of the Gammatone filterbank; additionally, based on a ratio between arithmetic and geometric power mean PNCC are able to filter some of the background noise. A much broader overview of auditory-based feature extraction methods is exposed by Stern and Morgan (2012).

Further physiologic and psychoacoustic research (Qiu et al., 2003) (Mesgarani et al., 2007) have shown the existence of neurons in the primary auditory cortex A1 of mammals specifically tuned to specific temporal or spectral modulations, and in some cases exhibit diagonal sensitivity patterns (such as vowel transients in speech). Spectro-temporal receptive fields (STRFs) are estimated patterns for time-frequency representations of stimuli optimally driving a neuron (or a group of neurons). To model such patterns, two-dimensional Gabor filters were consequently developed to model patterns observed in STRFs (Qiu et al., 2003), owing to the localized spectro-temporal patterns explicitly coded in A1. Kleinschmidt and Gelbart (2002) investigated if a set of those psychoacoustically parametrized filters, could extract meaningful information for robust ASR.

A challenge when designing filters for ASR is to determine a set of suitable parameters to produce a robust feature set able to deal with environmental noise, low signal-to-noise ratios, reverberation or even channel distortions. Schädler et al. (2011) proposed a Gabor filterbank based on specific physiologically-motivated temporal and spectral modulation frequencies, which resulted in relative improvements of the word error rate (WER) by 30 – 45% compared to a MFCC baseline for ASR (Meyer et al., 2012), and 21% for speaker identification (Lei et al., 2012).

In similar studies, a multitude of Gabor filters were employed to cover a wide range of modulation frequencies, and parsed as input to a large number of neural nets for merging the feature streams (Zhao and Morgan, 2008). Ezzat et al. (2007), based on 2D discrete cosine transforms, extracted spectro-temporal information to transform time-frequency patches of a spectrogram.

Previously, we explored the applicability of Gabor filters arranged in a filterbank as input to DNN-HMM back-end on the Aurora 4 task, which resulted in relative improvements of almost 20% over standardized filterbank features and 60% over MFCC results (Castro Martinez et al., 2014). Meanwhile, Chang and Morgan (2014), using a different convolutional neural network initialized with a different set of Gabor filters, obtained fruitful results on the same task as well as in a re-noised version of wall street journal. Subsequently, Baby and van Hamme (2015) proposed yet another auditory-based feature extraction method, which consist in low-frequency amplitude modulated spectrograms computed from low-passed-filtered half-way rectified signals; together with a DNN-HMM back-end, obtaining very similar WERs as we did for Aurora 4 and 19.6% phone error rate on the TIMIT corpus.

These studies support the idea of merging auditory-based features with deep learning architectures to get the best of both worlds, however, the cause behind the gain of such combination is still unknown. We believe the benefit comes from these representations which provide the deep learning decoder with more discriminable cues for the speech recognition task. Our aim with this paper is to validate this hypothesis by lowering the baseline word error rates (WER) in three different recognition tasks (Aurora 4, CHiME 2 and CHiME 3) and assess the discriminability of the features encoded by Gabor filters. We pursue the latter objective analyzing the activations obtained from the DNN through a robust metric of separability in feature space. The

indicated measure is the similarity between classes; being those the clustered context dependent triphone HMM states mapping to the same phoneme.

The remainder of this paper is structured as follows: we describe in detail the Gabor filterbank, along with the baseline features, the setup of the deep neural network and the criteria used for the analysis in the methods section. Results are presented in the following section, then a brief discussion depicting the results and the paper conclusions afterwards.

## 2. Methods

In this section we describe the auditory-based ASR features, i.e., Gabor features and the baseline filterbank features, the speech corpora, as well as the hybrid classification system which comprises a deep neural network and hidden Markov models. In the final part of this section, we present a method to assess the relevance of the auditory-based input streams for deep learning in comparison to the baseline features.

### 2.1. Gabor Filterbank features

Inspired by observations in spectro-temporal receptive fields in the auditory cortex (cf. previous section), we used a set of two-dimensional Gabor filters arranged in a filterbank to extract ASR features from speech signals. The procedure, depicted in Fig. 1, consists of three stages:

Initially, logarithmic Mel-spectrograms were extracted from the speech signals following the ETSI Distributed Speech Recognition Standard (201 108 v1.1.3 2003) with the only difference of using 31 frequency channels instead of 23. For the signals with a sampling frequency of 16 kHz, it provides a similar frequency resolution as the common 23 channels for 8 kHz data employed, for instance, in the Aurora 2 task (Hirsch and Pearce, 2000). Log-Mel-spectrograms were chosen as a starting point because they approximate the logarithmic compression of amplitudes and the non-linear frequency mapping of the auditory system. In the second stage, the spectrograms were convolved with every 2D filter in a modified version of the Gabor filterbank from (Schädler et al., 2011).

A Gabor filter is the product of a complex sinusoid function (1) and traditionally a Gaussian window; we replaced the latter with a Hann window (2) to obtain better recognition scores due to better modulation frequency characteristics, as reported in (Meyer et al., 2012). The periodicity of the carrier sinusoid was defined by the radian frequencies  $\omega_n$  and  $\omega_k$  ( $n$  and  $k$  denoting time and frequency index, respectively), which allowed the Gabor filters to be tuned to particular spectro-temporal directions, as well as purely temporal ( $\omega_k = 0$ ) or purely spectral ( $\omega_n = 0$ ) modulations.

The number of oscillations for the localized filters was kept constant for all filters, with a value of 3.5 as suggested by Schädler et al. (2011). This procedure is similar to wavelet processing and would result in infinitely large filters for modulation frequencies of zero; hence, all filters were limited to a maximum size (in this case 69 frequency channels and 99 time frames). The envelope width was parametrized by the window

lengths  $W_n$  and  $W_k$  and the center frequency channel  $k_0$  and center time frame  $n_0$ .

$$s(n, k) = \exp(i\omega_n(n - n_0) + \omega_k(k - k_0)) \quad (1)$$

$$h(n, k) = \frac{1}{4} \left[ 1 - \cos\left(\frac{2\pi(n - n_0)}{W_n + 1}\right) \right] \left[ 1 - \cos\left(\frac{2\pi(k - k_0)}{W_k + 1}\right) \right] \quad (2)$$

The Gabor filterbank contains a set of temporal, spectral and spectro-temporal filters to cover a wide range of modulation frequencies. Because frequency mapping is approximately linear at frequencies below 800 Hz rather than strictly logarithmic, spectral modulation frequencies are expressed in cycles per channel<sup>2</sup>. The specific modulation frequencies were chosen so that the transfer functions of the filters exhibit a constant overlap in the modulation frequency domain.

To account for modulations arising from syllable structure in spoken language, temporal modulations of 2.4 and 3.9 Hz were included as in the slightly modified filterbank presented in (Meyer et al., 2011) in addition to higher modulations also considered by Schädler et al. (2011). This resulted in 59 pairs of spectral and temporal modulation frequencies.

With 59 spectro-temporal filters and 31 frequency channels, the resulting feature vectors would have been rather high-dimensional (1829 components). However, filters with a large spectral extent produce highly correlated output between adjacent channels hence relatively small changes in the feature values when shifted by one frequency channel. Therefore, many frequency channels of larger filters were discarded from the feature matrix, while all channels were conserved for filters with the smallest spectral extent. This was achieved by choosing the channel centered on 1 kHz (which should contain information relevant for speech recognition) as well as channels obtained by shifting the current filter by one fourth of its spectral size and preserving its center frequency channel. Furthermore, as the Mel-spectrogram spectral size is smaller than the biggest Gabor filters, zero-padding was implemented to match the spectral content for the 2D convolution and to preserve the same number of features per frame without introducing significant boundary effects, the initial and last frame columns were padded on both temporal ends respectively.

The shifting value was selected based on the minimum window overlap needed for a perfect reconstruction of the spectrogram according to Nyquist-Shannon theorem. Alternative methods such as LDA and PCA were analyzed in Schädler et al. (2011). Critical sampling is designed to discard only redundant information, thus the number of selected channels lied between 1 (for  $\omega_k = 0$  cycles/channel) and 31 ( $\omega_k = \pm 0.25$  cycles/channel), and the feature dimension was reduced to 657.<sup>3</sup>

<sup>2</sup>Another unit could be cycles per mel, we opted for cycles per channel because it takes into account the mel scaling and also placing mel-frequency channels into bins. The mel-definition used in this work comes from the ETSI implementation calculated as follows:  $Mel(x) = 2595 \log_{10}\left(1 + \frac{x}{700}\right)$

<sup>3</sup>The original code used for feature extraction can be found in this repository: <https://github.com/m-r-s/reference-feature-extraction>

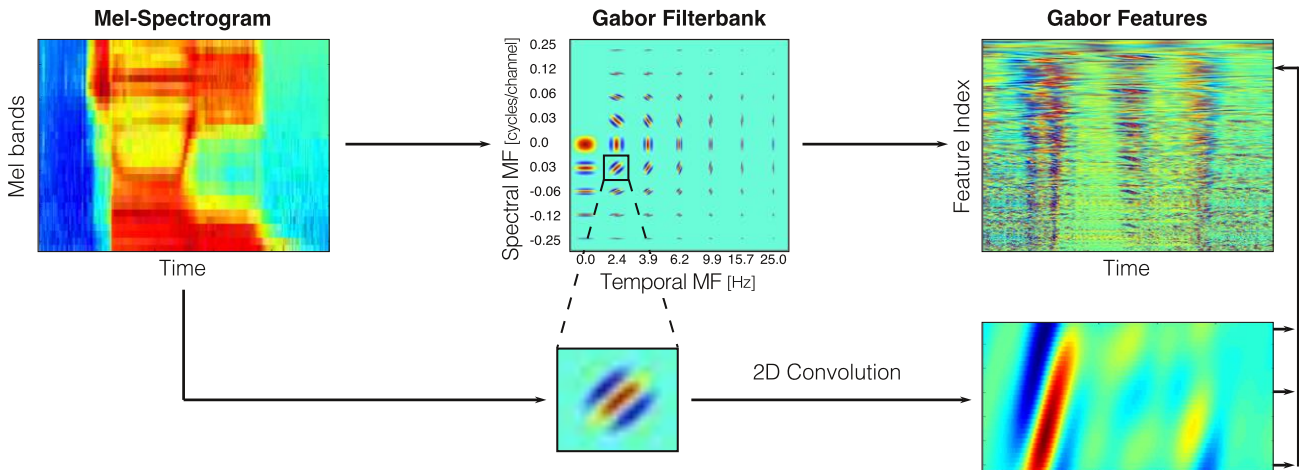


Figure 1: Gabor filterbank feature extraction procedure. The input log Mel-spectrogram is convolved by each of the 59 filters in the filterbank starting from the top and from left to right; the one taken as example below shows the contribution of this particular filter, the output is critically sampled and concatenated vertically giving the final 657-dimensional the complete feature vector representation.

## 2.2. Gabor Filterbank Subgroups

Given the wide range of spectral and temporal modulation frequencies taken into account in the Gabor filterbank, we became interested in knowing which particular set of filters is most relevant for DNN-based speech recognition. Hence, we divided the original filters into sets with low, medium and high temporal modulation frequencies, which resulting features are referred to as LTM (derived from filters with temporal modulations of 2.4 and 3.9 Hz), MTM (6.2 and 9.9 Hz) and HTM (15.7 and 25 Hz), respectively. Because critical sampling only removes spectral channels, all 3 subgroups are left with exactly the same channels.

Earlier experiments using a GMM-HMM recognizer and the Gabor filterbank indicated each individual 2D filter contributes to the noise robustness observed on the Aurora 2 task (Schädler et al., 2011), so we expected subgroups to perform worse than the full dataset. The filters selected from the complete set are highlighted in Fig. 2.

There were mainly two reasons for choosing this type of subdivision: first, our spectral modulations are given in cycles/channel and thus are more difficult to interpret and compare results than our temporal modulations in Hz. Secondly, lower temporal modulation frequencies have consistently been remarked as being most important in speech perception and recognition in the literature, which can be re-evaluated with this subdivision. It is important to mention how selecting a particular center frequency as temporal modulation does not necessarily exclude (only attenuates) the rest of the frequencies in the spectrum as the Gabor filters are broadband.

Kanedera et al. (1998, 1999) concluded the most useful linguistic information come from modulation frequencies components in the range of 2 to 16 Hz (with 4 Hz as predominant component), and components above or below this range could degrade recognition accuracy. Drullman et al. (1994a,b) measured the perception of speech synthesized with several temporal

envelopes for each frequency band and established the interval of modulation spectrum components between 4 Hz and 16 Hz to be the most critical for speech intelligibility; additionally, they reported a marginal contribution of modulation frequencies above 16 Hz when the lower components were present. Furthermore, the feature extraction procedure developed by Tchorz and Kollmeier (1999) performed the best on temporal modulations around 6 Hz. All these studies provide strong reasons to expect LTM to outperform HTM.

Owing to the critical sampling, the feature vectors extracted with the Gabor filterbank do not contain an equal amount of channels from each filter, i.e. the number of bands produced by the convolution of the log-Mel-spectrogram with each filter depends on the size of the filter, however, as each aforementioned subgroup includes all the spectral modulation frequencies, the dimension of output is 202 for every subgroup.

## 2.3. Baseline

Raw Mel-Filterbank features have been found to outperform MFCC features in recognizers with DNN-based architectures and hence serve as baseline features (Mohamed et al., 2012): Log-mel-spectral coefficients (MFSC) are obtained from the 31 channels of the same spectrogram used for calculating the Gabor features. As these filterbank representations include more information of the original Mel-spectrogram than MFCCs (where only the first 13 bands are selected), a DNN is less constrained to create any structure of the input data; thus provides us a better contrast for our "hand-crafted" features.

A conventional triphone GMM-HMM recognizer was built prior to the deep neural network in order to obtain target labels via force-alignment. Per speaker a single feature-space maximum likelihood linear regression transform was calculated to train this model adaptively.

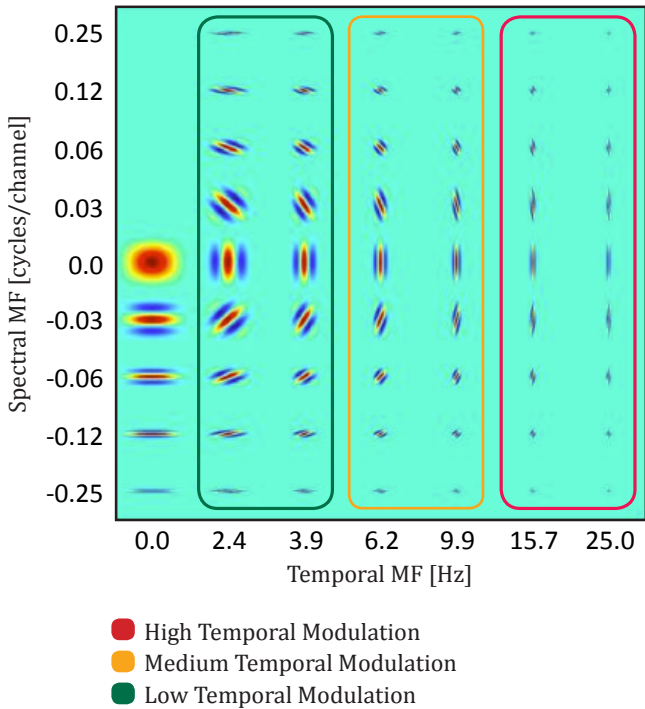


Figure 2: Gabor filterbank subgroups.

#### 2.4. DNN implementation

The deep neural network (DNN) was based on the one described in (Vesel et al., 2013). Recognition experiments were conducted using the Kaldi ASR toolkit (Povey et al., 2011). Due to the capability of unsupervised pre-training using Restricted Boltzmann Machines (RBM), which provides a deep hierarchical representation of the training data, we opted for Karel’s recipe.

In principle, this implementation can be summarized in two phases: pre-training and cross-entropy tuning. On the former phase, a stack of RBMs, also known as a deep belief network (DBN) (Mohamed et al., 2012), was trained in a greedy fashion one layer at a time using contrastive divergence as described by Hinton (2010).

On the latter phase, serving as a backbone for the final network, the DBN was fine-tuned to classify frames into triphone-states using an independent (development) set and the cross entropy between the network output and the labels as a cost function.

The training was done in up to 20 epochs (stopping when the relative improvement was lower than 0.001). The starting learning rate was 0.008 (halving it every time the relative improvement was lower than 0.01) and no momentum nor regularization techniques (such as  $L1$  &  $L2$ ) were applied. A soft-max layer of approximately 2000 units was attached to the end of the DNN to output the most likely posterior probabilities of each context-dependent HMM state.

The resulting DNN had 2048 sigmoid neurons on each of the six hidden layers. Optimization via stochastic gradient descent was performed on a graphics processing units for speeding pur-

poses. The size of the input layer varied depending on the type of the feature, for any given one 11 frames are spliced to provide a context of  $\pm 5$  frames.

In a nutshell, for every feature a GMM system was trained without changing any baseline configurations (except for the feature themselves) to provide the alignment of context dependent states to frames; then pre-training is performed to initialize the DNN, which uses the class labels provided by the GMM system; after fine-tuning the DNN is retrained, only this time it uses the labels produced by the DNN instead.

#### 2.5. Discriminability Criteria

Usually ASR Systems are evaluated in terms of the word error rate over a testing set. As we wanted to have a better understanding of the particular relationship between deep learning and auditory-based features we decided to observe the activations from the DNN output layer instead of analyzing the features separately.

Phoneme discriminability characterizes the performance of the learned representations even if the DNN is trained to deliver scaled probabilities of the senone HMM states, because each transition can be seen as a branch of a correspondent decision tree. The roots of those trees are the central phoneme of the trained triphones and are used to create phoneme classes from the clustered branches.

We selected a list of phonemes in the ARPA format (ARPA-BET) and gathered a group of activations corresponding to only the frames labeled as the phonemes in this list. The corpora used for this analysis had a disparate number of examples for each phone, so we created separate lists, one for the large vocabulary tasks (i.e., Aurora 4 and CHiME 3) and a different one for CHiME 2.

Labels were taken from the clean sets when possible to ensure they convey the spoken message; for Aurora 4 using the given clean close-microphone condition; for CHiME 3, as there are no available clean labels, we used the ones produced by the force-alignment from the best performing setup; the CHiME 2 clean labels were used from forced alignment system detailed in (Kabir et al., 2010).

Owing to the high dimensionality of the activations, a measure of separability robust to reparametrization was needed. As a criterion for assessing how well a particular feature separates the input into distinguishable classes we chose the cosine similarity. For being  $L_2$  based, this metric is invariant to rotation of the coordinate system and thus allowed us to compare the discriminability among the different features. The cosine similarity is defined as:

$$S(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \|\vec{v}_2\|} \quad (3)$$

Where each vector is the centroid of all the gathered examples for a given phoneme class (mean and variance normalized), the numerator is the inner product between the correspondent phoneme classes and the denominator is the product of their norm. The cosine similarity measures the relationship between

two vectors represented by a value between 0 and 1; the closer this value gets to 0 the wider the angle formed by these two vectors.

Generally, the similarity represents how close the phoneme manifolds are projected in the hyperspace, therefore a higher value increases the difficulty for phoneme separation performed by the DNN. Classes with a larger distance are less likely to be confused, conversely smaller angles (similarity values close to 1) lead to higher misclassifications. By calculating the similarities between every phoneme in the list with each other, we obtained the similarity matrices shown in the following section.

These matrices would ideally be identity matrices, so the more a similarity matrix resembles an identity matrix the better the classification capabilities of the system. For each corpus a similarity matrix was calculated to get a better understanding of the relevance of the information encoded by the auditory-based features in combination with the DNN for the recognition task.

## 2.6. Corpora

We performed a series of speech recognition experiments using different Corpora to prevent the system configuration and post-analysis from being adapted to a particular task. The Aurora 4 Corpus and the one used for the 3rd CHiME (Computational Hearing in Multisource Environments) Challenge are both derived from the same large vocabulary continuous speech recognition task, namely Wall Street Journal (Garofalo et al., 2007); the latter, however, also contains real recordings in noisy environments which provides more realistic data for the analysis.

The small vocabulary GRID-based corpus originated from the 2nd CHiME Challenge was of particular interest for us because the test set is available in multiple signal-to-noise ratios (SNRs) and therefore enables an analysis across different noise levels, furthermore, being a short vocabulary task allow to detect if there is a particular effect from the vocabulary size or the language model.

### 2.6.1. Aurora4

The Aurora 4 framework (Parihar et al., 2004) was used to assess the impact of additive noise from different sources and the effect of channel distortions; it is a large vocabulary continuous speech recognition task derived from the standard LDC Wall Street Journal (WSJ0) corpus. We opted for the multicondition set for training, which consists of 7137 utterances from 83 independent speakers, one half of the 16 kHz files were recorded with the close-talk Sennheiser HMD-414 microphone, the other half using one of 18 different types of microphones.

Each half was further subdivided; no noise was added to one fourth (893 utterances) while the remaining three-fourths (2676 utterances) were corrupted with one of six different types of noise (car, babble, restaurant, street, airport and train) at randomly selected SNR conditions between 10 and 20 dB. The test set included in the framework was extracted from the WSJ0 5,000 word closed-vocabulary task which consists of 330 utterances from 8 speakers repeated in the same 14 conditions used in the train set at 5 to 15 dB SNR.

### 2.6.2. CHiME3

For the third CHiME Challenge (Barker et al., 2015), sentences contained in the WSJ0 corpus were recorded using a 6-microphone tablet in four *real-life* scenarios: caf, street junction, public transport and pedestrian area. Additionally, the task includes also *simulated* noisy utterances to assess the value of generated data, as it is easier and cheaper to obtain and could be potentially useful for training purposes.

A total of 12 US English speakers (6 male and 6 female) ranging in age from 20 to 50 years were recorded after short test sessions to ensure each speaker performed the reading task correctly. An interface showed the talkers approximately 100 sentences to be read; those were recorded in an isolated booth (which served as basis for the simulated data) and in each location as described above. The training set comprised 1600 real noisy utterances (4 speakers x 100 sentences x 4 scenarios), whereas the development and test set consist of the same 410 and 330 utterances from the WSJ0 corpus, randomly divided in 4 subgroups and read on each scenario, resulting in 1640 and 1320 utterances respectively.

We used only the "noisy" set (utterances from the frontal closest microphone Channel 5) for training and testing to exclude from the analysis uncontrolled gains as a result of the speech-enhancement technique.

### 2.6.3. CHiME2-GRID

The CHiME2-GRID dataset (Vincent et al., 2013) from the second CHiME challenge was included for two main reasons: to have an estimate of the class separation performance from auditory based features and deep learning over different SNR levels and to verify if the proposed set up performed well on small vocabulary tasks. The GRID corpus (Cooke et al., 2006), from which this data was extracted, consist of 6-word sequences read by 34 speakers of the form: <command:4> <color:4> <prepos.:4> <letter:25> <digit:10> <adverb:4>, (e.g. "bin green at C 5 now") where the numbers in brackets indicate the number of choices per word.

These utterances were generated using binaural noise recordings from a head and torso simulator in a living room and mixed with the GRID data at 6 different SNR levels from -6 dB to 9 dB in steps of 3 dB. Moreover, each utterance was convolved with a set of head impulse responses simulating speaker movements and reverberation to make the task more realistic. We used the isolated noisy 16 kHz 500 utterances from each of 34 speakers as the training set, and the 600 utterances at each of the 6 SNR conditions as test sets.

## 3. Results

We experimentally confirm the effectiveness of auditory-based Gabor features across three different speech recognition tasks. The performance in terms of word error rate (WER) for four Gabor feature sets (full filterbank and the three subgroups according to the temporal modulation frequencies) and the filterbank baseline is presented in Table 1

For simplicity the 14 conditions from Aurora 4 were grouped into 4 subsets: "A" and "B" correspond to the clean and noisy

recorded using the close-talk microphone, respectively, and likewise "C" and "D" for the clean and noisy utterances with different channel characteristics introduced by the different secondary microphones. The CHiME 3 rows come from the real-recordings (real) and the simulated data (simu) parts of the test set. The bottom rows are the WER from the CHiME 2 test set, the number in parentheses indicates the corresponding SNR.

The features obtained from Gabor filters with low temporal modulations (LTM) produce notably the worse results in all three tasks. Conversely, the representations derived from filters from high temporal modulations (HTM) perform consistently better than all the others in each task. The second to best features for all conditions, except Aurora 4 "A", are the ones extracted from medium temporal modulation Gabor filters (MTM), whereas the complete Gabor filterbank (Gabor) produced features robust in noisy environments but not as good as the ones from raw filterbank (MFSC) in cleaner scenarios.

	MFSC	Gabor	LTM	MTM	HTM
Aurora 4 (A)	3.9	3.9	12.4	4.6	3.0
Aurora 4 (B)	7.5	8.5	21.2	7.8	5.6
Aurora 4 (C)	12.3	8.8	20.4	9.0	6.3
Aurora 4 (D)	22.2	19.2	34.6	18.4	15.4
CHiME 3 (simu)	23.3	30.8	41.6	20.2	15.2
CHiME 3 (real)	36.0	40.1	50.8	26.6	21.0
CHiME 2 (9dB)	4.8	6.1	7.3	4.9	4.4
CHiME 2 (0dB)	12.6	12.3	15.4	11.9	10.7
CHiME 2 (-6dB)	26.0	20.3	25.3	20.6	19.9

Table 1: Word Error Rates for the three speech recognition tasks comparing the baseline and the auditory-based Gabor features processed with Deep Neural Networks

The whole Gabor filterbank yields similar results as MFSC on conditions "A" and "B" of the Aurora 4 framework; on conditions "C" and "D" the effect of different channel characteristics can be appreciated as the WER decreases drastically even in the absence of additive noise. The WER increase from condition "C" to "D" is almost uniform (9 – 10%) for all features except LTM and is considerably larger than the one from "A" to "B".

Concerning the CHiME 3 task, the WER difference between the *real* and the *simulated* test sets can be used to quantify the generalization capabilities of the systems trained on different features because the generated data do not entirely capture the acoustic complexity of a *real-life* scenario. For MFSC features this difference is the biggest.

For the small vocabulary task, auditory-based Gabor features show robustness against low SNRs (except LTM); for Gabor and MTM this advantage is revealed when lowering the SNR below 0 dB, whereas HTM yield lower WERs even at positive levels. MTM perform almost the same as MFSC at 9 dB and 0 dB. The actual assignment evaluated in the first track of the CHiME 2 Challenge was to correctly recognize the letter and

digit tokens. To relate the recognition scores to the post-analysis shown in Fig. 4, where more phoneme samples were needed, we based our WER scores on the whole utterances.

To test whether the low temporal modulation filters are the culprits for the higher WER, we decided to restrict network-related optimization effects with the following configurations on Aurora 4: a) LTM + HTM (referred to as LHTM); b) MTM + HTM (MHTM); c) features produced by DC filters + HTM (DCHTM), the former being the filters with temporal modulation frequency of 0; d) random noise + HTM (RHTM) and e) a matrix of 0's + HTM (ZHTM). The last two set-ups contain either uniformly distributed numbers in the range of  $[-1, 1]$  or the an equal number of zeros, both matching the size of HTM, thus making all configurations but the third one equidimensional. Average WERs are shown in Table 2.

	LHTM	MHTM	DCHTM	RHTM	ZHTM	HTM
Aurora 4	13.10	11.55	11.35	11.02	10.75	9.66

Table 2: Average Word Error Rates for the Aurora 4 task comparing HTM and 5 additional configurations: LTM + HTM (LHTM), MTM + HTM (MHTM), features produced by DC filters + HTM (DCHTM), random noise + HTM (RHTM) and a matrix of 0's + HTM (ZHTM)

HTM outperformed the remaining configurations. Adding LTM drastically increases the error, more so than MTM, features from DC filters, zeros or even 0's. Random noise is slightly more detrimental than 0's as a complement to HTM; likewise MHTM performs worse than DCHTM. Only average results are shown as the same trend is observed consistently over each condition shown in Table 1.

Owing to the performance of HTM, a post-analysis was implemented using the discriminability criterion described in section 2.5. Fig. 3 shows the similarity matrices of MFSC and HTM on the Aurora 4 task. In order to reduce the bandwidth of the matrices, the phonemes were arranged through an implementation of the sparse reverse Cuthill-McKee ordering (Gilbert et al., 1992). For readability a threshold of  $\cos(45^\circ)$  was set in order to filter out the angles wider than  $45^\circ$ ; thus the non-zero elements represent similarity values above 0.7, which are likely candidates for phoneme confusions. A clear distinction is remarked between consonants on the upper left corner and vowels on the lower right. Both features produced almost the same similarity patterns for vowels, in this corner, the most noticeable confusions are between phones: /EY/ & /IY/, /AA/ & /AO/, /AE/ & /EH/ and /L/ & /OW/.

For consonants MFSC produced more confusions than HTM. The highest similarities between 2 phonemes occur among cases: /T/ & /D/ for both features; the values on cases /P/ & /B/, /DH/ & /B/, /T/ & /K/, /P/ & /K/ and /M/ & /N/ are predominantly higher for MFSC. A pattern of multiple confusions appears on the MFSC figure for phones: /V/, /D/ and /DH/ with respect to /T/, /K/, /N/, /P/ and /B/; whereas on HTM figure the same pattern has almost vanished (except for /D/ & /T/ and /DH/ & /B/). Phonemes /P/ & /F/ yield almost the same similarity for both features.

Fig. 4 shows the similarity matrices of MFSC on the left

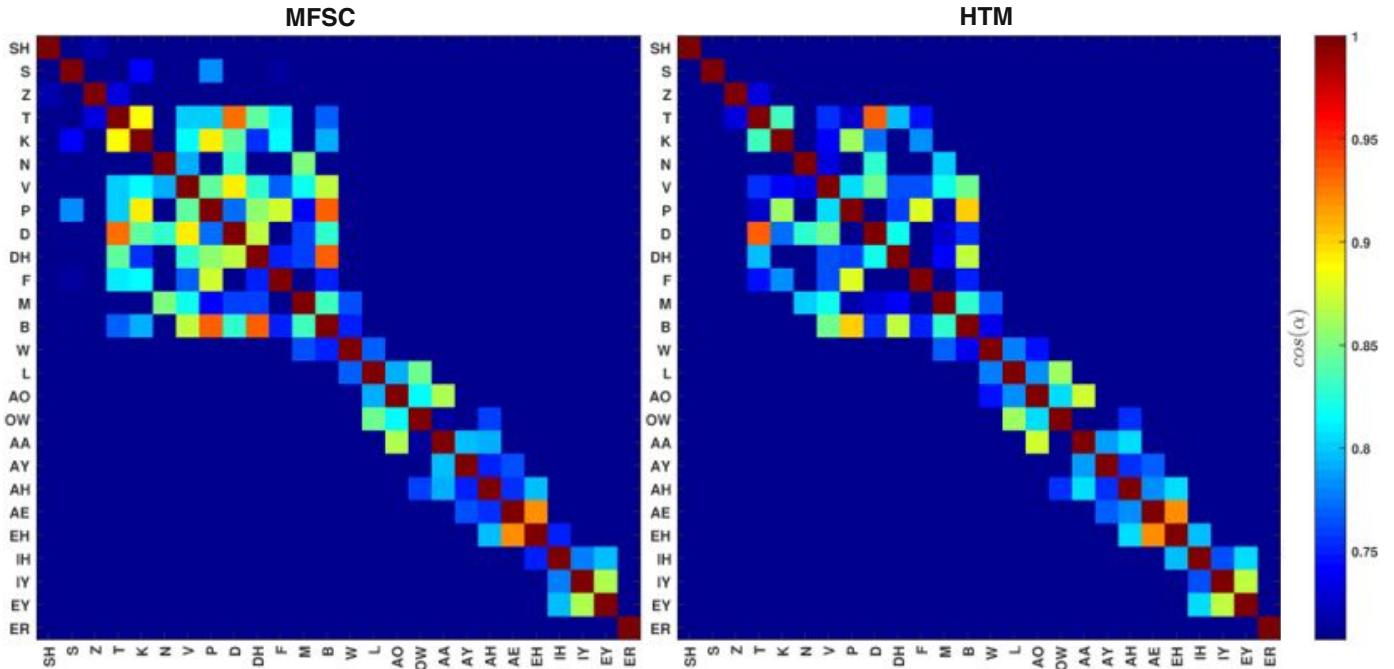


Figure 3: Similarity matrices of MFSC (left) and HTM (right) from the Aurora 4 corpus. For readability a lower threshold was set to  $\cos(45^\circ)$  so that angles wider than  $45^\circ$  would be ignored.

side and HTM on the right side calculated from the CHiME 2 framework. The upper row containing figures (a) and (b) were computed using the test set at 9 dB SNR and  $-6$  dB on lower row with figures (c) and (d). This time the threshold was set at  $60^\circ$  as the number of phonemes were reduced, thus the matrix elements represent values above the similarity value of 0.5. Nevertheless, we still refer to elements whose similarity value is above 0.7 as confusions.

For MFSC the similarity between the phoneme classes /Z/ & /S/ is high even at 9 dB and greater than the one of HTM. There is also a strong similarity of phonemes /B/ & /G/ for both features, but in Fig. 4 (c) the value gets closer to 1. With respect to phoneme /F/ there are high similarities with phonemes /B/ and /P/ for MFSC features whereas these values are below threshold in Fig. 4 (b) and almost so in Fig. 4 (d) for HTM.

At  $-6$  dB SNR several confusions appear on MFSC among the pairs of phonemes /R/ & /L/, /IH/ & /UW/ and /IH/ & /R/ and on HTM between the phonemes /R/ & /W/. In Fig. 4 (c) a *noisy* pattern, comparable to the one formed on phonemes /V/, /D/ and /DH/ in Fig. 3 for MFSC, can be observed forming on phoneme /F/ with respect to almost every phoneme from /N/ to /S/ (except phoneme /L/); this pattern is once again diminished for HTM (Fig. 4 (d)).

Another noticeable likely confusion among the four graphs is the one between /B/ & /P/; for both conditions (i.e. 9 dB and  $-6$  dB), however, the similarity increased twice as much for MFSC from 0.86 to 0.96 compared to HTM in which case the value went from 0.80 to 0.85. A similar pattern occurred in phonemes /IY/ & /UW/ and /T/ & /S/. Conversely, between phonemes /IY/ & /IH/ and /N/ & /AW/ the same deterioration of approximately 0.05 units was observed on both features when the SNR lowered 15 dB.

To further explain the robustness of HTM an extra set of experiments was conducted: firstly, for the most challenging scenarios in Aurora 4 (for instance, using a secondary microphone in a train station additive noise), both the features and the activations were analyzed to inspect saliency of features for specific speech sounds and to review the resulting performance in terms of phoneme classification on basis of the posteriorgram.

We found high temporal modulations produced a more confident decision (based on the activation strength) of the accurate label. Fig. 5 exemplifies this analysis condensing the evaluation performed on the activations under the worst performing scenarios. To analyze the structure of the projected classes from a trained setup, we recurred to the visualization technique called t-distributed stochastic neighbor embedding (t-SNE) proposed by Van der Maaten and Hinton (2008), which allowed us to observe the distribution of the target classes in low-dimensional manifolds.

Secondly, we gathered main confusion patterns from the activations produced by all features in the CHiME 2 task. Fig. 6 summarizes the most relevant confusion patterns on HTM and MFSC over all SNR conditions of the test set. Overall, the similarity value decreases when the SNR increases which supports the idea of more separable projections leads to better recognition scores.

Among the phonemes studied, the most confusable patterns were /B/ & /P/, closely followed by /Z/ & /S/ for MFSC and /G/ & /B/ for HTM, peaking at a highest similarity of 0.83 for HTM and 0.91 for MFSC corresponding to an estimated minimal separation between classes of approximately  $33^\circ$  and  $24^\circ$  respectively. The following pairs seem to be equally challenging for both features: /T/ & /S/ and to a lesser extent, /G/ & /B/ and /G/ & /P/. The pair of phonemes /V/ & /T/ is clearly more



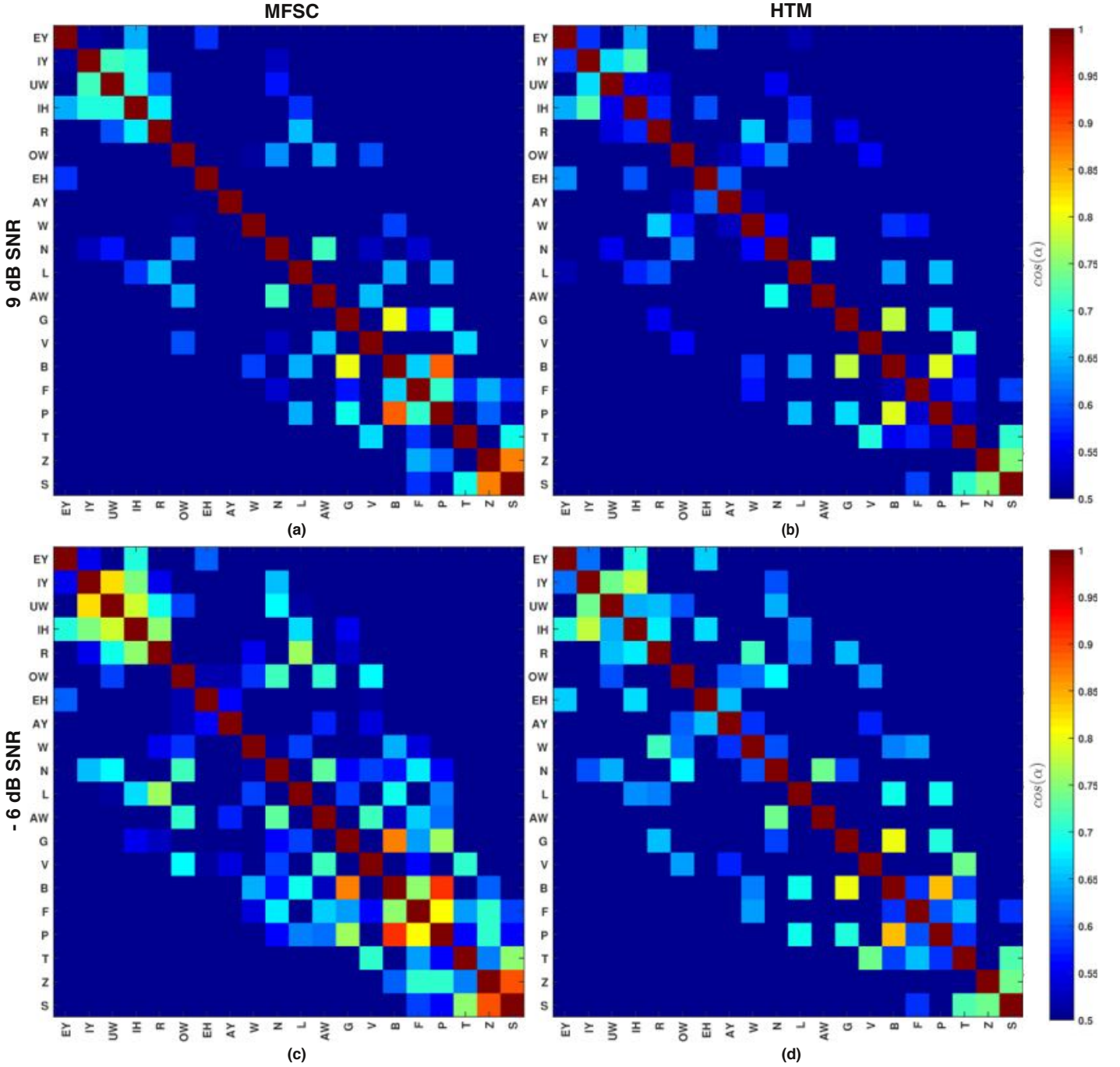


Figure 4: Similarity matrices of MFSC (left) and HTM (right) from the CHiME 2 Corpus. On the upper row the SNR is 9 dB and  $-6$  dB on the lower row. For readability a lower threshold was set to  $\cos(60^\circ)$  so that angles wider than  $60^\circ$  would be ignored.

discernible for MFSC; the opposite trend is observed in the case of /B/ & /P/, /Z/ & /S/, /B/ & /F/ and /F/ & /P/.

#### 4. Discussion

The results presented in Table 1 clearly show how exploiting a particular set of Gabor filters with high temporal modulation frequencies in combination with a deep neural network provides a boost in performance on three different recognition tasks. HTM yielded the lowest error rates, even in clean conditions where the baseline MFSC achieve already a high accuracy.

Meyer and Kollmeier (2010) used a stochastic approach to determine Gabor filter parameters relevant for ASR and found positive contributions for a wide range of temporal modulation frequencies from 2 to 25 Hz. Similarly, RASTA-PLP features, a range of 2.6–20 Hz was found to be useful. Hence, we expected the subgroups with a limited modulation range to perform worse than the complete Gabor filterbank, especially because fully connected DNNs are capable of handling correlated signals more effectively than traditional models. The results, however, show otherwise; both MTM and HTM outperform the whole

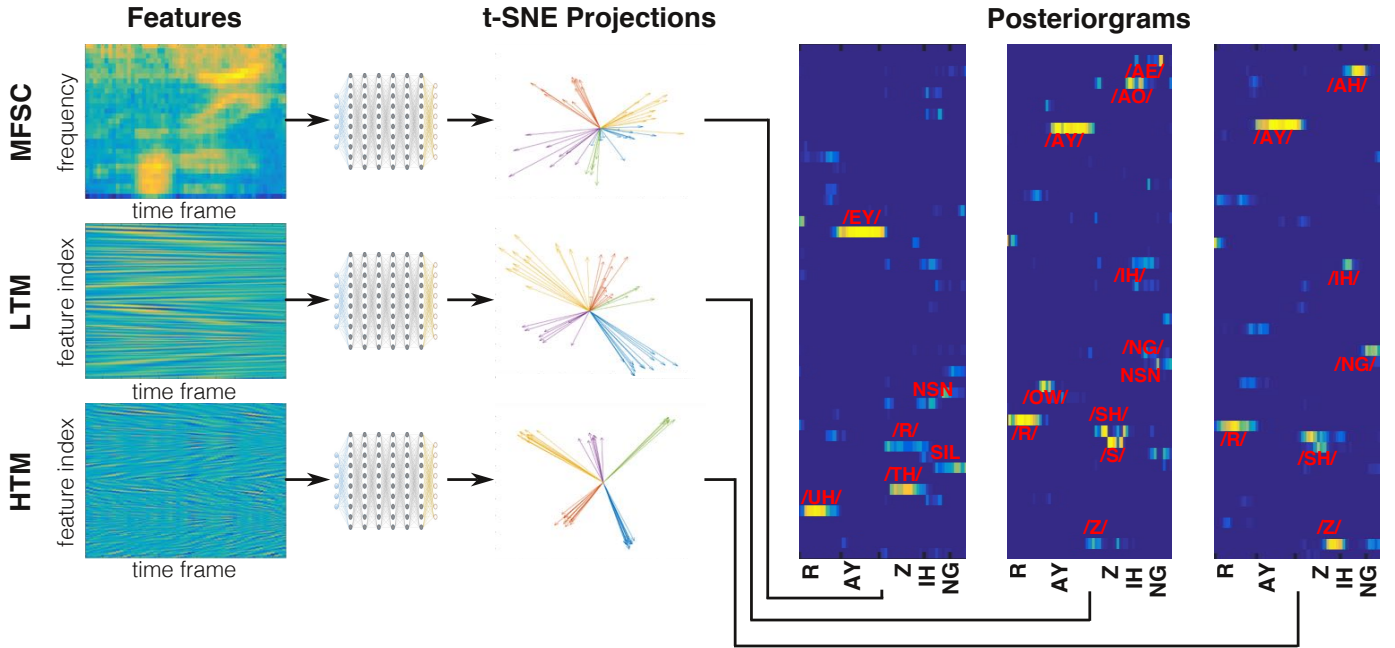


Figure 5: Example of the word *raising* extracted from one of the most challenging Aurora 4 scenarios (restaurant noise with secondary microphone). The first column are the features of this 53-frame segment, the t-SNE projections in the middle representing the discriminability per frame in the 5 phoneme classes, finally the resulting posteriorgram highlighting the strength of the activations produced by the DNN.

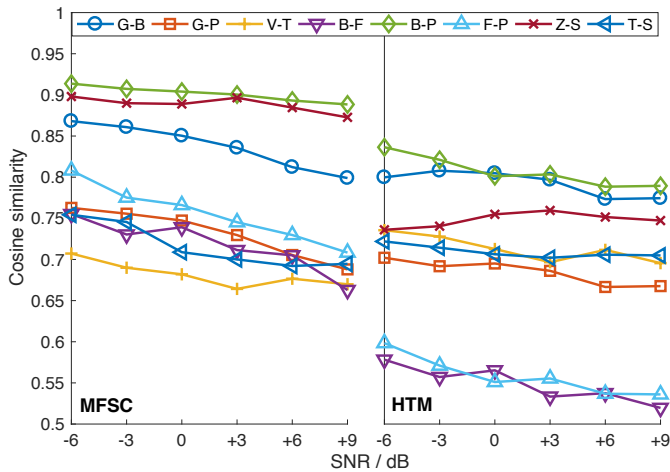


Figure 6: Confusion Patterns for the most relevant pairs of phonemes from the CHiME2 task for MFSC and HTM at different SNR conditions.

filterbank.

Moreover, the inclusion of lower temporal modulation frequencies (2 – 4 Hz) seems to severely harm the representations extracted by the Gabor filterbank as LTM features yield the highest WER in all conditions and corpora. This is in contrast to previous studies Kanedera et al. (1998, 1999), Drullman et al. (1994a,b) and Tchorz and Kollmeier (1999) that indicate high temporal modulation frequencies (above 16 Hz) to deteriorate performance on speech related tasks, from perception, recognition and intelligibility. Ganapathy and Omar (2014) arrived to a similar conclusion, their band-pass filtering results suggest 15 Hz is an optimal upper cut-off limit for speech recognition

performance in noisy conditions.

Kanedera et al. (1998) found 4 Hz to be the dominant component encoding the most useful linguistic information. In contrast to these contributions, the proposed high temporal modulations for Gabor processing do not filter out sharply contiguous regions among the spectra; by design the Gabor filterbank contain constant-Q filters, therefore their bandwidth is proportional to the center modulation frequency.

In other words the higher the modulation frequency the broader the bandwidth. So even both subgroups of filters reach contiguous frequencies, the gain of the individual transfer function is higher in neighboring frequencies for filters with higher central modulation frequency<sup>4</sup>.

We argue one of the reasons focusing on 16 and 25 Hz as a center frequency boosts recognition performance is because the produced features mimic the important strategy found in human listening to rely on localized patterns in the time-frequency representation (*glimpsing*), as pointed out in (Cooke, 2006), where the target signal dominates the noise, hence speech-relevant information can be extracted from the glimpses encoded in HTM.

The size of filters presumably also plays an important role: Given that low temporal modulation filters have the largest temporal extent in the Gabor filterbank, they produce stronger temporal smearing which might prevent the DNN from extracting phoneme-specific patterns. Neurons in the first layer compute a linear function of the input; feeding them with shorter segments of the spectrogram (i.e. smaller filters) allows higher layers containing non-linearities to learn more sparse and distributed

<sup>4</sup>For more details about the individual gain of each filter we refer the reader to (Schädler et al., 2011)

features, thus resulting in fewer confusions as discussed below.

Recently, Chait et al. (2015) conducted experiments to support the idea of multi-time resolution processing taking place in human speech perception. Their findings disprove low temporal modulations as sufficient for speech recognition and remark the need of a model to include higher modulation frequencies as well. Such a combination has been tested (on a GMM-HMM recognizer) by Hermansky and Fousek (2005) extending the RASTA processing (mentioned in Section 1) with a bank of two-dimensional filters to incorporate temporal trajectories of critical-band spectrograms; this step is followed by a TANDEM feature extraction. Albeit an akin approach to feature extraction with the complete Gabor filterbank; in this work, we found higher temporal modulations alone the most beneficial for ASR.

Each recognition task evaluated different aspects, for instance, Aurora 4 highlights the effect of additive noise and channel distortions at positive SNR levels. The changes in WER between conditions "A" and "B" as well as "C" and "D" represent the effect of additive noise. In the same way the detriment from "A" to "C" and "B" to "D" measures the effect of different channel characteristics. The complete Gabor set appears to be more robust against channel distortions compared to MFSC with consistent improvements for test sets C and D. This robustness is preserved for HTM plus a higher robustness against additive noise with an average relative improvement of 29% is achieved over the MFSC baseline.

Among the features shown in Table 2, HTM has the lowest number of feature components and yielded the lowest WER, suggesting there is not a significant effect in the recognition scores due to dimensionality. Concatenating random noise to HTM decreases performance slightly more than just adding zeros which reflects the capacity of the network to ignore uninformative input.

Surprisingly, the combination of HTM and the features extracted from DC filters resulted into lower WER than when combining MTM and HTM, even though the former was the second best feature from the ones compare in Table 1. We assume the MHTM combination contains more redundant components than DCHTM, which should have a detrimental effect.

Several successful approaches have been reported on the Aurora 4 task focusing on improved speech enhancement or feature extraction in deep learning systems: For instance, an exemplar-based speech enhancement proposed by Baby and van Hamme (2015) resulted in a WER of 11.9%. Chang and Morgan (2014) investigated Gabor filters in convolutional deep neural networks and obtained a 16.6% WER.

Similarly, improved net architectures have been investigated: (Rennie et al., 2014) trained an order statistic network with an *annealed* version of the dropout regularization method obtaining 10.0% WER. Geiger et al. (2014) achieved a 13.3% WER by implementing a long short-term memory recurrent neural network (for its ability to exploit temporal context) in combination with a non-negative matrix factorization for speech enhancement. Mitra et al. (2014) worked on both approaches and switched the fully connected deep neural network for one with convolutional layers together with vocal tract length normalization and lowered the WER on all conditions using a uniformly weighted

combination of 5 acoustic features (WER: 14.1%).

With a relatively simple approach of replacing the feature extraction, a WER of 9.7% was obtained in this study, which potentially could be further reduced by combining it with the above-mentioned methods (especially regarding more elaborate net architectures and regularization methods).

The CHiME 3 corpus focuses on the application of speech recognition technologies in *real-world* scenarios and sets a benchmark for comparing the value of artificially generated data for training and testing purposes. Among the features tested, HTM yielded the lowest WER difference between the real and simulated test data, which indicates an improved generalization when combining DNNs with these features. Additionally, because both Aurora 4 and CHiME 3 tasks are based on WSJ and share some acoustic scenarios, conditions "D" (noisy, different microphone characteristics) from the former should be comparable with the (simu) condition from the latter. This is not the case for LTM and the whole Gabor filterbank which suggests the non-shared noises could particularly harm LTM and thus the whole Gabor filterbank as well.

Concerning the third CHiME challenge itself, every system in the top 10 made several substantial changes to the baseline including augmentation of training data, speech enhancement, denoising, feature extraction, and improving or replacing the acoustic and language models. The top-ranked (Yoshioka et al., 2015) included a pre-processing model based on spectral masking and beamforming; however, the deep learning architectures were trained on MFSC features. Vu et al. (2015) focused mainly on speech enhancement via non-negative matrix factorization and beamforming as well and also replaced the HMM decoder for a recurrent neural network.

From the previous challenge, Moritz et al. (2013) indicated the modules developed in a hearing research environment were compatible and provided an incremental gain when combined; therefore the aforementioned systems could potentially benefit from the inclusion of auditory-based features as we show in this work.

Owing to the relatively low grammatical complexity in the first track of the CHiME 2 Challenge, the contribution of the language model can be delimited, thus the WER depends more on the acoustic model trained on the features we want to compare. In this task, the error rates at the lowest SNR highlights the robustness of auditory-based features. Given that our error rates include complete utterances (not exclusively the letters and digits tokens), they are not directly comparable to the studies submitted to the challenge.

To get a better understanding of how relevant is the information provided to the DNN, the similarity analysis was conducted. It revealed HTM are able to better separate phoneme classes, potentially resulting in fewer confusions during classification. To support the results obtained from this analysis, we recurred to acoustic properties defined by Jakobson and Halle (1956) (referred as distinctive features).

Using these binary properties most confusions can be explained based on spectral properties of the phoneme classes; for instance, vowels and approximants confusions are not sufficiently covered in Jakobson/s distinctive features scheme. Even

if there are articulatory and perceptual properties can be used to describe these classes, some confusion patterns observed on Fig. 4 (c) such as: /IH/ & /UW/, /IY/ & /UW/ and /IH/ & /R/ are unexpected as the phonemes involved are acoustically far apart from each other.

The similarity analysis on the Aurora4 corpus (shown in Fig. 3) exposed the discrimination capabilities of baseline MFSC and HTM. Because the distribution of vowels in the multidimensional space produce similar patterns for both features, we focused on consonant confusions to explain the improvements obtained with HTM, starting with the high-similarity pair /T/ & /D/. Both phonemes share many acoustic properties, except the former has longer duration, reduced voice onset time, and higher total amount of energy with greater spread across the spectrum (typical characteristics of the distinctive feature known as *tense*) and the latter, being *voiced*, presents periodic low frequency excitation.

The same distinction applies to the /P/ & /B/ confusion, additionally, both phonemes have a energy in the lower frequencies (property denominated *grave*) but only the /B/ presents energy on the closure phase, hence a steep transition is formed from the occlusion to the burst. This spectral change is enhanced by spectro-temporal features, thus we believe it is the main factor for the low number of confusions with HTM.

In terms of acoustic properties, /K/ differs from /T/ in being *compact* and *grave*; the first property refers to the concentration of energy in a particular region of the spectrum. Once again, the smaller confusion from HTM could be due to an accurate detection of the spectral transition from burst to aspiration in frequencies below 2 kHz. Between /P/ & /K/, the key distinction is the diffuse spectrum (opposed to *compact*) of the former; therefore, this pair of phonemes is spectrally more similar than /T/ & /K/ and thus has a higher similarity value.

The consistent confusion pattern observed among the plosives with /V/, /D/ and /DH/ presumably arises from their shared property *voiced*. The periodic low frequency excitation, spectral tilt and burst frequency of stop consonants is severely deteriorated by additive noise even at medium signal-to-noise ratios. MFSC features encode the energy from the frequency bands so this effect is particularly harmful to these features.

Finally, the similarity analysis on the CHiME 2 allowed us to directly observe which confusions appear by decreasing the SNR. Note the analysis does not include the CHiME 3 corpus because there is no clean data for the real recordings, which hinders generating high-quality phoneme labels with forced alignment as well as the calculation of the phoneme cluster statistics. In spite of sharing several phoneme confusions with Aurora 4, there are some others to consider: /Z/ & /S/ share almost all distinctive features except /Z/ is *voiced*, this property can be adequately encoded by filters with a high spectral modulation. /B/ & /G/ are voiced stops and their spectra bear some resemblance, although, in the case of /G/ it is *compact*.

The phoneme /T/ is *discontinuous* meaning there is an abrupt spectral transition, whereas /S/ is *strident* presenting high energy noise dominated by high frequencies. While conceptually HTM could detect the abrupt transition of /T/, this property is mostly unnoticeable in isolation because it shows before the phoneme

is pronounced. For the /Y/ & /IH/ confusion, the former is *tense*, which is a difficult distinctive feature to represent by either feature (also shown in Fig. 3 for the /T/ & /D/ confusion), as a longer temporal context might be needed and could account for appearing in all conditions.

In Fig. 4(c), a consistent confusion pattern occurs for phoneme /F/ with respect to almost every phoneme from /N/ to /S/ (except phoneme /L/); this *noisy* pattern is due to the negative the *discontinuous* property, which is presumably why this pattern is not present in Fig. 4 (d). The confusion pair /P/ & /F/ is noticeable in all conditions despite being as well distinctively *discontinuous*. We think a possible reason is that in some cases, for individual frames of 25ms duration, the spectrum of the frication phase in /P/ resembles the one of a short /F/; for some other cases the sudden transient from the closure to the frication phase or from the frication to the aspiration phase of the /P/ is better detected by HTM, hence the lower similarity value.

## 5. Conclusions

The present study assessed the contribution of Gabor features in combination with deep learning architectures. We found a subgroup of filters within the Gabor filterbank capable of reducing even further the word error rates in the three different recognition tasks (Aurora 4, CHiME 2 and CHiME 3). The proposed HTM outperformed the MFSC baseline. These features are capable of detecting quick spectro-temporal transitions within 40 ms time windows and exhibited robustness against channel distortions, low signal-to-noise ratios and acoustically challenging *real-life* scenarios; they also perform better on clean-conditions.

Because the gains presented in Table 1 come from a relatively simple feature exchange, i.e. no additional speech enhancement, dereverberation or denoising techniques are applied, we assume it is straightforward to further improve the performance by combining one or more approaches including alternative deep learning approaches such as more elaborate net architectures and regularization methods.

The discriminability of MFSC and HTM was evaluated through the similarity analysis and explained in terms of distinctive spectral properties. The most relevant findings can be summarized as follows: Phonemes characterized as *grave*, *discontinuous* or *compact* exhibit spectro-temporal transients, hence they are less likely to be confused by HTM features. *voiced* consonants create consistent confusion patterns for MFSC features. *tense* phonemes are equally hard to distinguish for both features. The confusions of obstruent consonants are more representative of the performance difference between HTM and MFSC features in the presence of additive noise and channel distortions. Overall HTM features produced a more separable distribution of phones.

Finally, in this study we show DNN-based speech recognizers trained with Gabor features, particularly the ones exclusively using high temporal modulation filters, yield lower error rates as these features enhance the discriminability between the target classes.

## Acknowledgment

This work was funded by the DFG (Cluster of Excellence 1077/1 Hearing4All (<http://hearing4all.eu>), and the SFB/TRR 31 "The Active Auditory System" (<http://www.sfb-trr31.uni-oldenburg.de/>)) and by Google via a Google faculty award to Hynek Hermansky.

The authors thank Jon Barker for providing the clean labels for the CHiME 2 Corpus and Mats Exter whose expert advice in phonetics and phonology was of great help for the discussion. We thank the reviewers for the time and effort invested in earlier drafts, whose helpful comments helped enrich and clarify this manuscript.

## References

- Baby, D., van Hamme, H., September 2015. Investigating modulation spectrogram features for deep neural network-based automatic speech recognition. In: Proc. INTERSPEECH. pp. 905--909.
- Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third chime speech separation and recognition challenge: dataset, task and baselines. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).
- Castro Martinez, A., Moritz, N., Meyer, B. T., September 2014. Should deep neural nets have ears? the role of auditory features in deep learning approaches. In: Proc. INTERSPEECH. pp. 2435--2439.
- Chait, M., Greenberg, S., Arai, T., Simon, J. Z., Poeppel, D., Chait, M., 2015. Multi-time resolution analysis of speech. Name: Frontiers in Neuroscience 9, 214.
- Chang, S., Morgan, N., September 2014. Robust cnn-based speech recognition with gabor filter kernels. In: Proc. INTERSPEECH. pp. 905--909.
- Cooke, M., 2006. A glimpsing model of speech perception in noise. The Journal of the Acoustical Society of America 119 (3), 1562--1573.
- Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120 (5), 2421--2424.
- Drullman, R., Festen, J. M., Plomp, R., 1994a. Effect of reducing slow temporal modulations on speech reception. The Journal of the Acoustical Society of America 95 (5), 2670--2680.
- Drullman, R., Festen, J. M., Plomp, R., 1994b. Effect of temporal envelope smearing on speech reception. The Journal of the Acoustical Society of America 95 (2), 1053--1064.
- Ezzat, T., Bouvrie, J., Poggio, T., 2007. Spectro-temporal analysis of speech using 2-d gabor filters. Proc INTERSPEECH.
- Ganapathy, S., Omar, M., 2014. Auditory motivated front-end for noisy speech using spectro-temporal modulation filtering. The Journal of the Acoustical Society of America 136 (5), EL343--EL349.
- Garofalo, J., Graff, D., Paul, D., Pallett, D., 2007. Csr-i (wsj0) complete. Linguistic Data Consortium, Philadelphia.
- Geiger, J. T., Gemmeke, J. F., Schuller, B., Rigoll, G., 2014. Investigating nmf speech enhancement for neural network based acoustic models. INTERSPEECH. ISCA.
- Gilbert, J. R., Moler, C., Schreiber, R., 1992. Sparse matrices in matlab: design and implementation. SIAM Journal on Matrix Analysis and Applications 13 (1), 333--356.
- He, X., Deng, L., Chou, W., 2008. Discriminative learning in sequential pattern recognition. Signal Processing Magazine, IEEE 25 (5), 14--36.
- Hermansky, H., 1990. Perceptual linear predictive (plp) analysis of speech. The Journal of the Acoustical Society of America 87 (4), 1738--1752.
- Hermansky, H., Fousek, P., 2005. Multi-resolution rasta filtering for tandem-based asr. In: Proceedings of Interspeech 2005.
- Hermansky, H., Morgan, N., 1994. Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on 2 (4), 578--589.
- Hinton, G., 2010. A practical guide to training restricted boltzmann machines. Momentum 9 (1), 926.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE 29 (6), 82--97.
- Hirsch, H. G., Pearce, D., 2000. The AURORA Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. In: Proc. Automatic Speech Recognition: Challenges for the new Millennium. pp. 29--32.
- Jakobson, R., Halle, M., 1956. Phonology and phonetics. Mouton & Co. Printers, The Hague.
- Kabir, A., Giurghi, M., Barker, J., 2010. Robust automatic transcription of english speech corpora. In: Communications (COMM), 2010 8th International Conference on. pp. 79--82.
- Kanedera, N., Arai, T., Hermansky, H., Pavel, M., 1999. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28 (1), 43--55.
- Kanedera, N., Hermansky, H., Arai, T., 1998. On properties of modulation spectrum for robust automatic speech recognition. ICASSP Proceedings 2 (1), 613--616.
- Kim, C., Stern, R. M., 2009. Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. In: Proc. INTERSPEECH. pp. 28--31.

- Kleinschmidt, M., Gelbart, D., 2002. Improving Word Accuracy with Gabor Feature Extraction. *Proc. INTERSPEECH*, 25--28.
- Lei, H., Meyer, B. T., Marghafari, N., 2012. Spectro-Temporal Features for Speaker Recognition. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.
- Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* 22 (4), 745--777.
- Lippmann, R., 1997. Speech recognition by machines and humans. *Speech Communications* 22 (1), 1--15.
- Mesgarani, N., Stephen, D., Shamma, S., 2007. Representation of phonemes in primary auditory cortex: how the brain analyzes speech. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.
- Meyer, B., Spille, C., Kollmeier, B., Morgan, N., 2012. Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition. *Proc. INTERSPEECH*.
- Meyer, B. T., Kollmeier, B., 2010. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication* 53 (5), 753--767.
- Meyer, B. T., Ravuri, S. V., Schädler, M. R., Morgan, N., 2011. Comparing different flavors of spectro-temporal features for ASR. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. pp. 1269--1272.
- Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., Graciarena, M., 2014. Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In: *Proc. of Interspeech*.
- Mohamed, A.-r., Hinton, G., Penn, G., 2012. Understanding how deep belief networks perform acoustic modelling. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 4273--4276.
- Mohamed, A. R., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., Picheny, M., 2011. Deep belief networks using discriminative features for phone recognition. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 5060--5063.
- Moritz, N., Schädler, M. R., Adiloglu, K., Meyer, B. T., Jrgens, T., Gerkmann, T., Kollmeier, B., Doclo, S., Goetze, S., 2013. Noise robust distant automatic speech recognition utilizing nmf based source separation and auditory feature extraction. *Proc. of CHiME*.
- Pan, J., Liu, C., Wang, Z., Hu, Y., Jiang, H., 2012. Investigation of deep neural networks (dnn) for large vocabulary continuous speech recognition: Why dnn surpasses gmms in acoustic modeling. In: *8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. pp. 301--305.
- Parihar, N., Picone, J., Pearce, D., Hirsch, H.-G., 2004. Performance analysis of the aurora large vocabulary baseline system. In: *Signal Processing Conference, 2004 12th European*. pp. 553--556.
- Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M., 1992. Complex sounds and auditory images. In: *Auditory physiology and perception, Proc. 9th International Symposium on Hearing*.
- Povey, D., 2005. Discriminative training for large vocabulary speech recognition. Ph.D. thesis, University of Cambridge.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., and K. Vesel, N. G., 2011. The kaldi speech recognition toolkit. *Proc. ASRU*.
- Qiu, A., Schreiner, C. E., Escabi, M. A., 2003. Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *Journal of Neurophysiology* 90 (1), 456--476.
- Rennie, S. J., Goel, V., Thomas, S., 2014. Annealed dropout training of deep networks. In: *Spoken Language Technology Workshop (SLT), 2014 IEEE*. pp. 159--164.
- Sainath, T. N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P., Mohamed, A., 2011. Making deep belief networks effective for large vocabulary continuous speech recognition. In: *Automatic Speech Recognition and Understanding (ASRU)*, IEEE. pp. 30--35.
- Schädler, M. R., Meyer, B. T., Kollmeier, B., 2011. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. Submitted to the *Journal of the Acoustical Society of America*.
- Scharenborg, O., 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communications*, 336--347.
- Seide, F., Li, G., Chen, X., Yu, D., 2011. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: *Automatic Speech Recognition and Understanding (ASRU)*, IEEE. pp. 24--29.
- Stern, R. M., Morgan, N., 2012. Features based on auditory physiology and perception. *Techniques for Noise Robustness in Automatic Speech Recognition*, 207--243.
- Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automatic speech recognition. *The Journal of the Acoustical Society of America* 106 (4), 2040--2050.
- Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9 (2579-2605), 85.
- Vesel, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence discriminative training of deep neural networks. *Proc. Interspeech*.
- Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The second chimespeech separation and recognition challenge: Datasets, tasks and baselines. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. pp. 126--130.
- Vu, T. T., Bigot, B., Chng, E. S., 2015. Speech enhancement using beamforming and non negative matrix factorization for robust speech recognition in the chime-3 challenge. In: *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*.
- Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Yu, M. F. C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., Nakatani, T., 2015. The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In: *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*.
- Zhao, S., Morgan, N., 2008. Multi-stream spectro-temporal features for robust speech recognition. *Proc. INTERSPEECH*.