INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

DNN-based automatic speech recognition as a model for human phoneme perception

Mats Exter[1], *Bernd T. Meyer*[2]

[1]Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
mats.exter@uni-oldenburg.de, bernd.t.meyer@jhu.edu

Abstract

In this paper, we test the applicability of state-of-the-art automatic speech recognition (ASR) to predict phoneme confusions in human listeners. Phoneme-specific response rates are obtained from ASR based on deep neural networks (DNNs) and from listening tests with six normal-hearing subjects. The measure for model quality is the correlation of phoneme recognition accuracies obtained in ASR and in human speech recognition (HSR). Various feature representations are used as input to the DNNs to explore their relation to overall ASR performance and model prediction power. Standard filterbank output and perceptual linear prediction (PLP) features result in best predictions, with correlation coefficients reaching $r = 0.9$.

Index Terms: speech recognition, phoneme perception, models of speech intelligibility

1. Introduction

The practical importance of adequate models of human speech intelligibility has long been recognized and led to the development of index-based measures such as the Articulation Index (AI), the Speech Intelligibility Index (SII) and the Speech Transmission Index (STI). These measures can be called macroscopic in the sense that they evaluate longer stretches of speech, such as complete utterances. More recently, several studies have addressed the challenge of modelling speech intelligibility on a microscopic level, i.e. on a phoneme-by-phoneme basis, using a variety of ASR techniques. Cooke [1], for instance, developed a so-called "glimpsing model" that uses fragments of spectro-temporal representations of speech in noise as input to a Hidden Markov Model (HMM) ASR backend. Jürgens et al. [2, 3], on the other hand, used the output of an elaborate psychoacoustic model as input to a dynamic time warping (DTW) backend. Finally, Marxer et al. [4] have presented a framework for the evaluation of such microscopic models.

There are several areas where microscopic speech intelligibility models can potentially be applied. Firstly, they can further our knowledge of the processes involved in speech perception. Secondly, they can be used in the development of speech coding and transmission algorithms and in the development and fitting of state-of-the-art hearing aids. Finally, they can provide guidance in the improvement of ASR systems.

In the present study, building on earlier work by Meyer and Kollmeier [5], four different feature types that are inspired by properties of the human auditory system to varying degrees were used as input to a DNN ASR backend. The resulting phoneme confusion matrices were analysed and compared with phoneme confusion patterns obtained from human subjects performing the same task. Two properties were investigated in detail: overall recognition performance and correlation between ASR and HSR performance. The latter can serve as a measure of microscopic model quality, a higher correlation corresponding to a greater similarity between ASR and HSR phoneme confusion patterns.

It should be noted that the model proposed in the present study is reference-free; i.e., neither are speech and noise separately known to the ASR system nor are any of the training utterances also used during testing.

2. Methods

2.1. Speech database

The speech database used in this study was the Oldenburg Logatome (OLLO) Corpus [6], which contains 150 different logatomes. Logatomes are simple phoneme triplets with identical outer phonemes, either consonant-vowel-consonant (CVC) or vowel-consonant-vowel (VCV). Each utterance was recorded three times with five different intrinsic variations. This was realized by asking the speakers to produce utterances with high and low speaking effort, high and low speaking rate, and rising pitch. Normal speaking style was also recorded as a reference condition.

The database contains recordings from 50 speakers (10 German speakers without regional dialect as well as speakers from several regions in Germany and Belgium with different dialects or accents) and a total of 133,403 utterances. The OLLO corpus is freely available for research in HSR and ASR. It can be downloaded from http://medi.uni-oldenburg.de/ollo/.

For the HSR listening experiments, a subset of the database was compiled; this selection contained utterances from four speakers without dialect (2 male, 2 female), spoken with six different intrinsic variations. The HSR test set thus contained 3,600 items (150 logatomes $\times$ 4 speakers $\times$ 6 intrinsic variations). For the ASR experiments, the same four speakers were used for testing, while the data from the remaining 46 speakers was used for training. The phoneme combinations recorded for the database are shown in Table 1.

2.2. Listening experiments

Six normal-hearing listeners (3 female, 3 male) between 18 and 35 years of age participated in the collection of the perceptual data. Participants were presented a series of logatomes in speech-weighted noise [7] at an SNR of -6.2 dB. Their task was to identify the central phoneme in the VCV and CVC

http://dx.doi.org/10.21437/Interspeech.2016-1285

	Central phoneme	Outer phoneme
Phonemes (VCV)	/b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /s/, /ʃ/, /t/, /v/, /ts/	/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/
Phonemes (CVC)	/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/, /aː/, /eː/, /iː/, /oː/, /uː/	/b/, /d/, /f/, /g/, /k/, /p/, /s/, /t/

Table 1: Overview of the phonemes contained in the Oldenburg Logatome Corpus. The initial and final phonemes were identical for all recorded logatomes. The combination of each central phoneme with each of the outer phonemes results in 150 utterances (70 VCV, 80 CVC).

logatomes, which were presented in random order. It should be noted that due to this experimental design, only substitution errors were possible (for the HSR as well as for the ASR experiments). Listening experiments were conducted using closed headphones (Sennheiser HDA 200) in a sound-insulated booth. The utterances were presented at a comfortable listening level (70 to 75 dB SPL for most of the listeners). In total, 21,600 responses were collected.

2.3. Feature types and ASR setup

Almost all ASR systems incorporate properties of the human auditory system in their choice of feature types to some extent; cf. [8] and [9] for reviews of the role that auditory representation can play in ASR. Four different feature types were used in the experiments: filterbank (FBank) features, mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP) features, and Gabor filterbank (GBFB) features; these feature types are in turn discussed below. In the present work, most of the feature extraction for the ASR as well as the ASR itself were performed with Kaldi [10].

The ASR experiments were performed at a number of different SNRs (again using speech-weighted noise [7]); training and testing for each experiment were matched in terms of SNR.

2.4. FBank, MFCC and PLP features

FBank, MFCC [11] and PLP [12] features are the most commonly used methods in feature extraction. FBank features are obtained by smoothing the short-time Fourier transform (STFT) magnitude, which is typically computed every 10 ms using an overlapping analysis window of 25 ms. They form the basis for the other feature types in this study, and they have been found to be a baseline that is superior to MFCCs when DNN-based classification is performed.

For the computation of MFCCs, pre-emphasis is applied to the signal before calculating the STFT. Each frame is then processed by a mel filterbank (which approximates the response of the human ear), compressed logarithmically and transformed to cepstral parameters by using an inverse discrete cosine transform. By selecting several (typically 12 or 13) lower cepstral coefficients, only the coarse spectral structure is retained. This processing results in mostly decorrelated features. In this study, delta and double-delta features are appended to form the final MFCC feature vector, resulting in 39-dimensional feature vectors.

PLP features incorporate further psychoacoustic constraints: Linear prediction coefficients are computed from a perceptually weighted, nonlinearly compressed power spectrum. The power spectrum is obtained with a bark filterbank with subsequent equal-loudness pre-emphasis and a compression based on Steven's power law (i.e., values are compressed by applying the cube root). The linear prediction coefficients are then transformed to cepstral coefficients.

2.5. Spectro-temporal Gabor features

Gabor features are calculated by processing log mel-spectrograms (i.e., sequences of FBank feature vectors over time) of the input signal with a set of 2-dimensional modulation filters. Filtering is performed by calculating the 2-dimensional convolution of the spectrogram and the respective filter. The time-aligned result of the convolution for all filters is used as a feature vector.

Gabor filters are defined as the product of a complex sinusoidal function $s(n,k)$ (with n and k denoting the time and frequency index, respectively) and an envelope function $h(n,k)$. In this notation, the complex sinusoid is defined as

$$s(n,k) = \exp\left[i\omega_n(n-n_0) + i\omega_k(k-k_0)\right],$$

and the Hann function that we chose as envelope (with the parameters W_n and W_k for the window length) is given by

$$h(n,k) = \left[0.5 - 0.5\cdot\cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right)\right] \cdot \left[0.5 - 0.5\cdot\cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right)\right]. \quad (1)$$

The periodicity of the carrier function is defined by the angular frequencies ω_k and ω_n, which allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including diagonal modulations. For this study, an arrangement in a filter bank [14, 15] was chosen due to the good results that were obtained in various speech tasks with this specific implementation [13, 16, 17]. 657-dimensionsal features were calculated in Matlab, using the same algorithm and parameter settings as in [13]. The filter set is shown in Fig. 1.

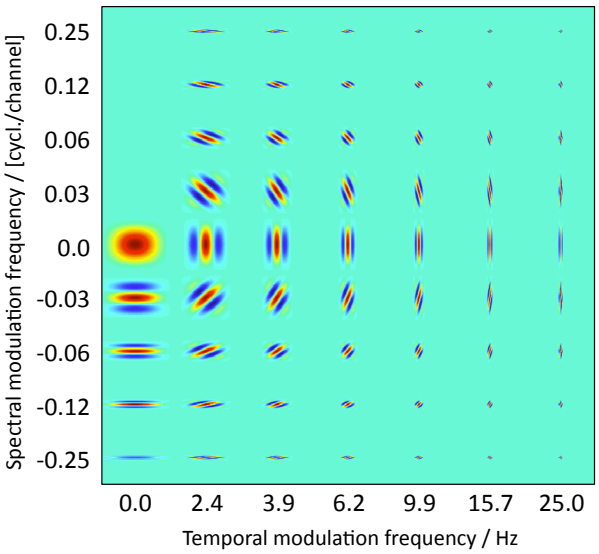

Figure 1: Real components of Gabor filters used for the filter bank, arranged by spectral temporal modulation frequencies.

2.6. DNN-based backend

All features were subsequently fed into a DNN/HMM classifier, using the Kaldi speech recognition toolkit [10]. The DNN had five hidden layers with 1024 units per layer. Input features were spliced with a temporal context of ± 5 frames and used as input to the DNN. A softmax transformation was applied to the DNN output.

Training of the DNN was performed in two steps. In the first step, an unsupervised pre-training of the network (functioning as a deep belief network, DBN) was performed to initialize its parameters, followed by a supervised fine-tuning of the network (now functioning as a standard feedforward DNN).

3. Results

In order to compare the features in terms of their recognition performance, phoneme recognition accuracies were calculated at SNR values between -39.2 dB SNR and 17.8 dB SNR at steps of 3 dB for all four feature types. The results are shown in Fig. 2.

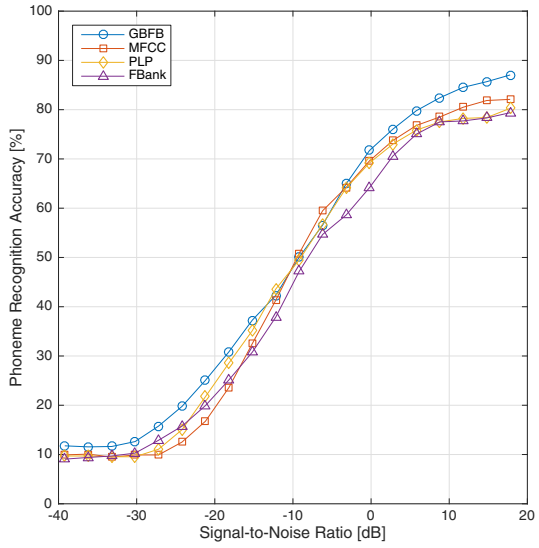

Figure 2: Phoneme recognition accuracy for different feature types in DNN-based ASR systems for the Oldenburg Logatome Corpus. (The HSR phoneme recognition accuracy at -6.2 dB SNR was 73.0%.)

Over most of the SNR values, GBFB features resulted in the highest recognition performance, followed by MFCC, PLP and FBank features. The HSR phoneme recognition accuracy at -6.2 dB SNR was at 73.0% still considerably higher than the best ASR performance at the same SNR.

As for the assessment of model quality, correlation coefficients were calculated between the phoneme recognition accuracies of ASR and HSR, again at the same set of SNR values for the ASR (HSR data was collected at a fixed SNR of -6.2 dB). The results are shown in Fig. 3.

A maximum correlation coefficient is observed at -6.2 dB SNR (for PLP features) or -9.2 dB SNR (for all other feature types). What is more, FBank and PLP features show the highest maximum correlation coefficient ($r = 0.89, p < 0.001$),

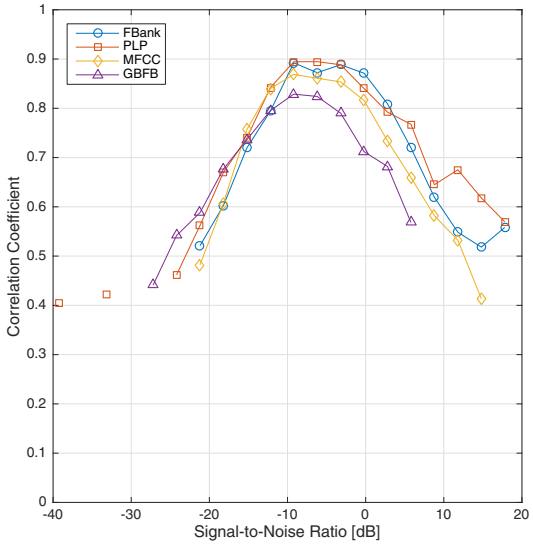

Figure 3: Model quality in terms of correlation coefficients between HSR and ASR (only significant values are shown). HSR data was collected at a fixed SNR (-6.2 dB). The plot shows prediction power of ASR phoneme recognition accuracies as a function on the SNR used for ASR (in training and test).

followed by MFCC ($r = 0.87, p < 0.001$) and GBFB features ($r = 0.83, p < 0.001$). Finally, at the very lowest SNRs, only PLP features show stable, significant results.

To assess the question of how the differences in the correlation coefficients between the feature types arise (or, to put it differently, where the source of any dissimilarities between ASR and HSR is), ASR phoneme recognition accuracies at -6.2 dB SNR are plotted against HSR phoneme recognition accuracies in Fig. 4.

At least at the given SNR, all four feature types show quite similar recognition patterns and correlations between ASR and HSR. This is in contrast to the findings in [5], where GBFB features departed markedly from the general pattern displayed by the other feature types.

4. Discussion

In the present study, auditory-inspired features were combined with DNNs in an attempt to model human speech intelligibility. Two measures were used for evaluation: phoneme recognition accuracy, and correlation between ASR and HSR.

As for the correlation as a measure of model quality, compared to the GMM/HMM-based approach in [5], the DNN-based approach in the present study has proven to be superior, since the SNR-mismatch (i.e. the mismatch between the SNR of the ASR where the correlation is at its maximum, and the SNR of the HSR) is far smaller or even non-existent (in [5], the SNR-mismatch was 6.2 dB for PLP features, while in the present study, it is 0 dB for PLP features and 3 dB for the other feature types). Also, the DNN-based approach is superior in terms of correlation values (in [5], the maximum correlation coefficient was $r = 0.84$, while here, it was $r = 0.89$, again for PLP as well as for FBank features). The perceptual training presented in [5] seems not to be required in the DNN context.

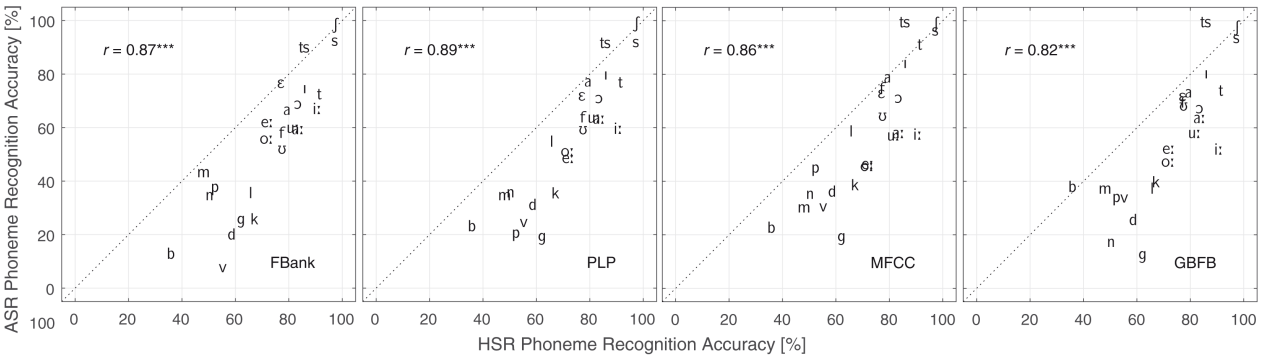

Figure 4: Comparison of ASR and HSR phoneme recognition accuracies at an SNR of -6.2 dB, obtained with different features. The feature types are given in the lower right, the correlation coefficients in the upper left corner of each subplot. The dotted diagonal lines represent lines through the origin with a slope of 1.

As for the phoneme recognition accuracy, GBFB features perform best (although not nearly as well as the HSR), but do not result in the best model prediction except for very low SNRs. This is in line with the findings of [5], where an even lower correlation of ASR and HSR recognition results was found when using GBFB features.

Generally speaking, a trade-off was found in the present work between recognition power on the one hand and model quality on the other hand: Over all SNRs, GBFB features showed the highest performance, followed by MFCC, PLP and FBank features; model quality (as measured by ASR–HSR correlation), on the other hand, was best for FBank and PLP features, followed by MFCC and GBFB features.

We think it is for two reasons interesting to note that GBFB features produce the least accurate phoneme predictions: First, they were proposed based on psychoacoustic and physiological evidence and hence should have a strong connection to auditory processing. Second, the best average performance in ASR is obtained with this feature type; hence, the model predictions should be less affected by the fact that ASR in general is often not on a par with human speech recognition (e.g., in extreme cases ASR is approaching chance performance while human listeners are still well above 50% phoneme error rate [18]).

This means that Gabor filterbank features exhibit a kind of processing strategy that seems quite different from the average healthy auditory system. This will be investigated in future research by dissecting the spectral, temporal, and spectro-temporal components to analyze if one of these filter groups is over-represented in the current feature design.

While the aim of the current study was to explore the applicability of the ASR system as a microscopic (i.e., phoneme-level) model of human speech intelligibility, the system could also be extended to predict sentence-level speech intelligibility by including a suitable language model.

5. Summary

In this paper, we have investigated to what extent an ASR system by itself can serve as a model of microscopic phoneme confusions in human listeners. Deep learning methods have had a very strong impact on speech research in the previous years and have been established as a standard in ASR. We have shown that using DNN-based classifiers can be beneficial for creating models of human speech perception as well, with high phoneme recognition accuracy correlation values between human and machine recognition, while an SNR shift between ASR and HSR test data is no longer required (in contrast to previous work). What is more, the resulting speech intelligibility model is reference-free, deriving its predictions from previously unknown, noisy data. Although the goals in ASR and models of HSR are different, we still found it surprising to observe that the best ASR features performed worst for the perception model and vice versa, reaching correlation values of $r = 0.9$.

6. Acknowledgements

This work was funded by the DFG (Research Unit FOR 1732 “Individualized Hearing Acoustics”, Cluster of Excellence 1077/1 “Hearing4all” and SFB/TRR 31 “The Active Auditory System”) and by Google via a Google faculty award to Hynek Hermansky (CLSP, Johns Hopkins University).

7. References

[1] Cooke, M. (2006). “A model of speech perception in noise,” J. Acoust. Soc. Am. 119, 1562–1573.

[2] Jürgens, T., Brand, T., and Kollmeier, B. (2007). “Modelling the human-machine gap in speech reception: Microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model,” in *Proceedings of Interspeech*, pp. 410–413.

[3] Jürgens, T., and Brand, T. (2009). “Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model,” J. Acoust. Soc. Am. 126, 2635–2648.

[4] Marxer, R., Cooke, M., and Barker, J. (2015). “A framework for the evaluation of microscopic intelligibility models,” in *Proceedings of Interspeech*, pp. 2558–2562.

[5] Meyer, B., and Kollmeier, B. (2010). “Learning from human errors: Prediction of phoneme confusions based on modified ASR training,” in *Proceedings of Interspeech*, pp. 1209–1212.

[6] Meyer, B. T., Jürgens, T., Wesker, T., Brand, T. and Kollmeier, B. (2010). “Human phoneme recognition as a function of speech-intrinsic variabilities,” J. Acoust. Soc. Am. 128, 3126–3141.

[7] Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001), “ICRA noises: Artificial noise

signals with speech-like spectral and temporal properties," Audiology 40, 148–157.

[8] Hermansky, H. (1998). "Should recognizers have ears?," Speech Commun. 25, 3–24.

[9] Stern, R. M., and Morgan, N. (2012). "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," IEEE Signal Process. Mag. 29, 34–43.

[10] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

[11] Davis, S., and Mermelstein, P. (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, 28, 357–366.

[12] Hermansky, H. (1990), "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. 87, 1738–1752.

[13] Castro Martinez, A. M., Moritz, N., and Meyer, B. T. (2014). "Should deep neural nets have ears? The role of auditory features in deep learning approaches," in *Proceedings of Interspeech*, pp. 2435–2439.

[14] Schädler, M. R., Meyer, B. T., and Kollmeier, B. (2012). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am. 131, 4134–4151.

[15] Meyer, B. T., Ravuri, S. R., Schädler, M. R., and Morgan, N. (2011). "Comparing different flavors of spectro-temporal features for ASR," in *Proceedings of Interspeech*, pp. 1269–1272.

[16] Tsai, T. J., and Morgan, N. (2012). "Longer features: They do a speech detector good," in *Proceedings of Interspeech*, pp. 1356–1359.

[17] Lei, H., Meyer, B., and Mirghafori, N. (2012). "Spectro-temporal Gabor features for speaker recognition," in *Proceedings of ICASSP*, pp. 4241–4244.

[18] Cooke, M., and Scharenborg, O. (2008). "The Interspeech 2008 consonant challenge," in *Proceedings of Interspeech*, pp. 1765–1768.