Learning from human errors: Prediction of phoneme confusions based on modified ASR training

Bernd T. Meyer and Birger Kollmeier

Medizinische Physik, Institute of Physics, University of Oldenburg

bernd.meyer@uni-oldenburg.de

Abstract

In an attempt to improve models of human perception, the recognition of phonemes in nonsense utterances was predicted with automatic speech recognition (ASR) in order to analyze its applicability for modeling human speech recognition (HSR) in noise. In the first experiments, several feature types are used as input for an ASR system; the resulting phoneme scores are compared to listening experiments using the same speech data. With conventional training, the highest correlation between predicted and measured recognition was observed for perceptual linear prediction features ($r = 0.84$). Secondly, a new training paradigm for ASR is proposed with the aim of improving the prediction of phoneme intelligibility. For this perceptual training, the original utterance labels are modified based on the confusions measured in HSR tests. The modified ASR training improved the overall prediction, with the best models ($r = 0.89$) exceeding those obtained with conventional training ($r = 0.80$).

Index Terms: automatic speech recognition, human speech perception, phoneme recognition

1. Introduction

A model that accurately predicts the speech intelligibility of normal-hearing and hearing-impaired listeners in noisy conditions may assist in the development and fitting of hearing aids, or for the development of speech coding algorithms. The majority of existing models uses knowledge about the human auditory system in order to predict either the long-term (or macroscopic) intelligibility (for example, the Speech Intelligibility Index (SII)), or the microscopic intelligibility, i.e., the intelligibility of short-term speech segments such as phonemes. Recently, several microscopic models were proposed that make use of techniques commonly applied to the problem of automatic speech recognition (ASR). For example, Jürgens and Brand [1] used a model that converts signals into a perceptual internal representation using an auditory model. Subsequently, a dynamic time warping algorithm was applied in order to successfully predict phoneme recognition rates in noise for normal-hearing listeners. Similarly, Cooke used a glimpsing model (that extracts the spectro-temporal cues exploited by listeners to detect noisy speech fragments) in combination with a standard ASR back end (Hidden Markov Model, HMM) for the prediction of consonant scores [2].

The present work differs from these approaches in that it uses the outcome of listening experiments to build a statistical model that predicts phoneme recognition scores (in contrast to using perceptual models of the human auditory system). As a first step, we investigate if the framework of current standard ASR systems is suitable for the prediction of human phoneme recognition rates, and which ASR front-ends for feature extraction should be preferred for modeling purposes. To this end, an ASR phoneme recognition task is performed, and several feature extraction schemes (both classic spectral feature coding and spectro-temporal features) are analyzed. The recognition scores are compared to results from human speech recognition (HSR), which were obtained with the same speech data used for ASR testing. Vowel and consonant recognition rates are considered for the comparison.

In the second set of experiments, a novel approach for ASR training is proposed: we train the classifier using the response of listeners from the HSR task to the presented utterance as new label for that utterance. This enables us to investigate if this approach results in more human-like errors, which would be beneficial when using ASR technology in applications of speech intelligibility prediction.

2. Experiments

2.1. Speech database and listening experiments

The speech database used in this study is the Oldenburg Logatome (OLLO) Corpus [3], 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 in 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, and a total of approximately 110,000 utterances. Ten German speakers without regional dialect (whose data is used in the present work), as well as speakers from different regions in Germany and Belgium with a dialect or accent have also been recorded. 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 listening experiments, a subset of the database was compiled; this selection contained utterances from four speakers without dialect (2M, 2F), spoken in six different intrinsic variations. The HSR test set contained 3,600 items (150 logatomes $\times$ 4 speakers $\times$ 6 intrinsic variations). The phoneme combinations recorded for the database are show in Table 1.

Six normal-hearing listeners (3F, 3M) 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 [4] at an SNR of -6 dB. Their task was to identify the central phoneme in the VCVs and CVCs, which were presented in random order. Listening experiments were

	Central phoneme	Outer phoneme
Phonemes VCVs	/b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /s/, /ʃ/, /t/, /v/, /ts/	/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/
Phonemes CVCs	/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/, /a:/, /e/, /i/, /o/, /u/	/b/, /d/, /f/, /g/, /k/, /p/, /s/, /ʃ/

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

conducted using closed headphones (Sennheiser HDA200) in a sound-insulated booth. The utterances were presented at a comfortable listening level (70-75 dB SPL for most of the listeners).

2.2. Exp. I: Feature types and ASR back-end

Four different feature types were used for the experiments. This includes the most common features for ASR, which encode the spectral envelope of short-term spectra of speech, i.e. Mel-frequency cepstral coefficients (MFCCs) [5], Perceptual linear prediction coefficients (PLPs) [6], and Rasta-PLPs [7]. The rastamat toolbox for Matlab was used to calculate the features [8]. For MFCC and PLP features, settings that reproduce the feature calculation of HTK [9] were chosen. For Rasta-PLPs, the model order of the linear predictor was set to 12, which resulted in 13-dimensional features per time step. The addition of delta and double delta features yielded 39-dimensional vectors for each feature type.

Additionally, Gabor features were extracted from the noisy data which are based on spectro-temporal filtering of mel spectrograms. The Gabor filter set was the same as the one employed in [10], i.e., 80 filters were used that captured purely spectral and temporal characteristics of the signal, as well as spectro-temporal cues. The 80-dimensional feature vectors were processed with a non-linear neural net (MLP) with 56 neurons in the output layer (corresponding to the 56 phoneme classes in the database used for the MLP training). The output of the neural net was decorrelated with a principal component analysis and used to train and test an HMM recognizer. The algorithm employed to find a suitable Gabor filter set, as well as the further processing stages are described in detail in [10].

ASR experiments were carried out with a Hidden Markov Model (HMM) with three states and five Gaussian mixtures per state. Noisy utterances with the same outer phoneme were used to train and test HMM implemented in HTK [9]. A setup was chosen in which only confusions between central phonemes occurred, resembling the identification of the central phoneme used as task in HSR. For this set of experiments, a training/test set was chosen which enables a comparison with the perceptual data: The test set contained the same 3,600 utterances (recorded from four speakers) which had been presented to human listeners. The training set contained data from six other speakers (3F, 3M) without regional dialect (13,829 utterances in total). This subdivision resulted in a speaker-independent ASR system and is therefore referred to as training/test *SI*.

Ideally, the same SNR should be employed in HSR and ASR to eliminate the effect of different energetic masking for both conditions. However, this often results in either a ceiling of HSR scores or in ASR performance close to chance. In order to determine if an SNR mismatch results in higher prediction performance, the ASR scores were obtained for several SNRs (-5, 0, 5, 10, 15, 20, 30 dB, matched SNR for training and testing) and compared to HSR scores (measured at a fixed SNR of -6 dB).

2.3. Exp. II: Modified ASR training

A modified ASR training was carried out with the aim of improving the prediction of HSR phoneme recognition scores (Fig. 1). The perceptual data obtained in listening experiments

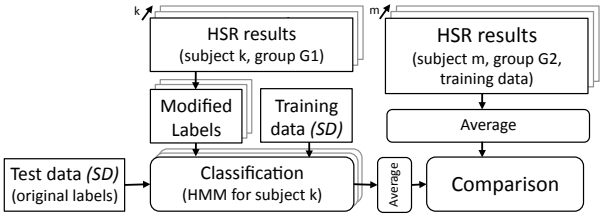

Figure 1: Scheme of modified ASR training. ASR labels are compiled based on the confusions measured in listening tests with human listeners.

was used to create label files that contain the mapping of single utterances to the individual responses of a listener. These label files replaced the original training label files.

An important requirement for any system used for the prediction of speech intelligibility in real-world applications is that a prediction can be made for *unknown* utterances (that have not been used during training). Since the HSR data was not available for the whole *SI* dataset, a different division into train and test set was required. The 3,600 utterances for which HSR data existed were equally split into a training and a test set. Since the amount of training data was limited, a speaker-dependent division was chosen, i.e., data of four speakers was contained in both sets. This set is referred to as training/test *SD*. The intrinsic variations and the number of representations for each logatome were equally distributed in both sets.

Furthermore, the HMM training requires specific labels for each utterance, i.e. an average over several subjects cannot be used as label. Hence, several HMMs were trained with the data from single listeners. We later use the average of those HMMs as model output. Finally, the listeners were randomly split into two groups. Data from the first group *G1* (1M, 2F listeners) served as input for the training procedure, while data from Group *G2* (1F, 2M listeners) was used for the evaluation of model prediction. This subdivision is necessary to avoid the prediction of data that has already been used during training.

3. Results

In order to assess the similarity of human and automatic phoneme classification, the correlation between HSR and ASR phoneme recognition scores was calculated. When ASR predictions based on perceptual training are reported, the average scores for Set *G2* are presented and compared to HSR data for the listener group *G1* (cf. Fig. 1). Additionally, the average phoneme recognition rates depending on the SNR for both experiments are presented.

3.1. Feature types

The phoneme recognition scores obtained with a speaker-independent recognizer are shown in Table 2. At high SNRs

Original Training, training/test set *SI*						
SNR	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB
MFCC	40.9	63.1	75.2	81.1	84.4	85.1
PLP	49.0	63.5	73.2	77.2	83.5	82.1
Rasta-PLP	49.1	65.2	71.5	75.1	84.0	85.1
MLP-Gabor	54.8	67.5	75.1	80.9	84.0	84.6

Table 2: Overall phoneme recognition rates in % for the speaker-independent ASR system with conventional training.

(5 dB or more), the best overall performance is obtained with MFCC features. For SNRs below 5 dB, the spectro-temporal Gabor features perform best. Human listeners achieved an average recognition rate of 75.5 %, which is equivalent to approximately half the error rate observed for ASR at low SNRs. The correlation of HSR and ASR phoneme recognition rates is shown in Fig. 2. The values were calculated for HSR at a fixed

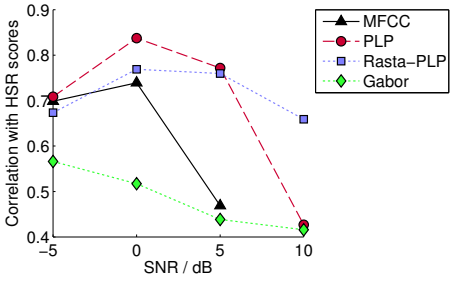

Figure 2: Correlation of ASR and HSR scores. Correlations for SNRs > 10 dB were not significant and are therefore not shown in the plot.

SNR (-6 dB) and ASR at multiple SNRs. The best model prediction was observed for PLP features with a mismatch of 6 dB between HSR and ASR ($r = 0.84$, $p < 0.0001$). The 6 dB mismatch also resulted in good model predictions for other feature types. Therefore, the phoneme confusions were analyzed in more detail for HSR at -6 dB SNR and ASR at 0 dB SNR (Fig. 3). While MFCCs, PLPs and Rasta-PLPs produce similar patterns and good correlations with HSR results, the Gabor features show a different pattern and a lower correlation.

Interestingly, the correlation between the scores of Gabor and MFCC features was rather low ($r = 0.56$, $p < 0.05$), while the overall recognition rates were comparable (cf. Table 2). Distinct differences between the feature types were for example observed for fricatives (/s/, /f/, /ts/, /ʃ/) which produced higher MFCC scores and for vowels for which Gabor features performed better. This shows that the differences between the feature types also extends to the recognition of specific phoneme classes, which could be exploited in ASR experiments by using a combination of both feature types.

3.2. Effect of modified ASR training

The recognition scores obtained with the *SD* training/test sets are presented in Table 3. To compare the results for Sets *SD* and *SI*, we first evaluated the performance with the *original* training (Table 3.a). The two main factors that differ between the ASR setup in Experiments I and II are the number of training utterances (with only 1,800 utterances in Exp. II compared to approx. 14,000 in Exp. I) and the speaker-dependent vs. speaker-

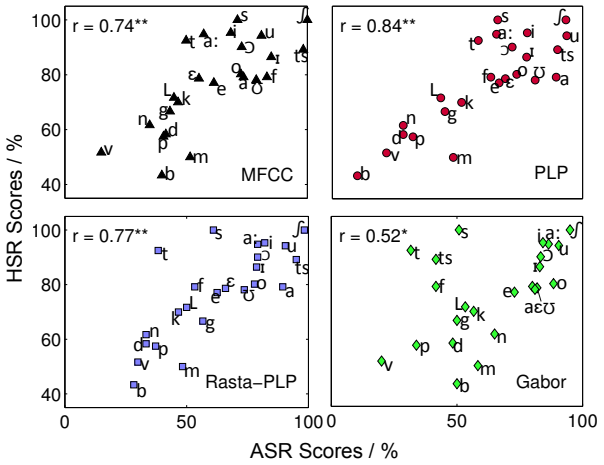

Figure 3: Comparison of HSR scores (SNR -6 dB) with ASR scores (SNR 0 dB) obtained with different features. The correlation coefficient is presented in the upper right corner of each subplot. Phoneme labels are given in IPA notation, where /l/ has been replaced with 'L' to avoid confusion with other phonemes.

a) Original Training, training/test set *SD*						
SNR	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB
MFCC	29.7	53.4	61.3	65.2	66.7	68.0
PLP	31.9	45.0	52.7	62.8	68.7	70.8
Rasta-PLP	29.2	44.4	51.4	53.3	55.2	62.9
Gabor	33.4	46.9	54.0	59.5	63.9	64.2

b) Perceptual Training, training/test set *SD*						
SNR	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB
MFCC	31.3	48.4	53.3	54.0	54.7	54.9
PLP	29.6	41.7	45.9	52.7	56.2	57.5
Rasta-PLP	28.0	40.1	44.2	45.3	45.3	49.8
Gabor	31.8	43.6	49.3	51.6	53.2	52.6

Table 3: Recognition scores in % for the speaker-dependent ASR systems with conventional training (a) and with perceptual training (b). Scores with perceptual training were obtained by averaging over six ASR systems, each trained with HSR data from a single listener.

independent task. The scores for Set *SD* are degraded by 18 % on average, which shows that simplification of the recognition task cannot compensate for the reduced amount of training material.

Not surprisingly, the modified training labels based on perceptual data resulted in a further degradation of recognition scores (8 % on average) throughout all SNRs and for all feature types (with the exception of MFCC features at -5 dB SNR). However, these strong degradations were not observed for the *correlations* between HSR and ASR scores: With the original training (Fig. 4.a) correlations were obtained that are comparable to the results shown in Fig. 2: The best model predictions were found for PLP and Rasta-PLP features, with a maximum value of $r = 0.80$. The perceptual training increases the best model prediction from 0.80 to 0.89 (PLPs at 10 dB SNR). As for Experiment I, Gabor features exhibited the lowest correlations for the majority of conditions.

When averaging over SNRs from -5 to +10 dB (at which

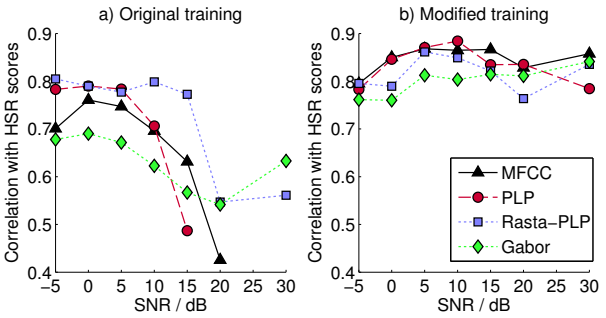

Figure 4: Correlation of ASR phoneme recognition rates with HSR results, depending on the SNR used for ASR training and testing. Only significant correlations are plotted.

relatively constant correlations were measured) and when taking all feature types into account, the perceptual training increases the correlations by 0.09. Furthermore, the perceptual training results in good phoneme predictions over a wider range of SNRs used in ASR, since significant r-values over 0.70 are obtained for all features at all SNRs that were considered.

4. Summary and conclusions

In this study, two experiments were presented which analyze (I) the applicability of different feature types for modeling human speech perception and (II) a novel approach for training ASR systems based on perceptual data from listening experiments.

The results from Exp. I show that good predictions of human phoneme recognition can be obtained with spectral features (MFCCs, PLPs and Rasta-PLPs) in combination with a standard back-end. This finding is in line with results from a study on consonant recognition in clean speech [12] that reported a high correlation of 0.81 between HSR and a standard ASR system based on MFCC features. Our results show that good predictions can be obtained even for noisy speech, and that PLP features seem to be better suited for modeling human speech perception than cepstral features. Therefore, this feature type should be preferred for modeling purposes. This result appears plausible since PLP features incorporate several findings of the human auditory system such as equal loudness weighting of the spectrum that are not covered by other feature types.

Additionally, it was shown that the best model predictions are obtained for a mismatch between the SNR used for the ASR model and the SNR employed for the human listening tests. This was attributed to the generally low performance of ASR at low SNRs (Table 2), which yields scores close to chance performance, thereby decreasing the correlation with human data. On the other hand, the effects of energetic masking cannot be modeled correctly at very high SNRs with ASR, since, for example, the masking of consonants with low energy is not covered by the model. The optimal compromise between these effects was found for an SNR mismatch of 6 dB between HSR and ASR model, which resulted in a highly significant correlation between PLP-based ASR and HSR of 0.84.

For Experiment II, a new speaker-dependent data set *SD* for training and testing was employed, which enabled the use of a training procedure based on data from listening experiments. This 'perceptual training' improved the ASR-based model predictions for all feature types and for all test conditions: While for conventional training the highest correlation with Set *SD* was 0.80, the perceptual training improved the correlation to 0.89 (Fig. 4), and therefore seems to be well-suited to predict human phoneme confusions in noise.

Although a considerable amount of HSR data was employed for the experiments (with a total 21,600 single presentations, or 3,600 presentations per listener), the model performance may be further improved using a larger amount of perceptual data. Specifically, a model that predicts phoneme confusions should preferably be speaker-*independent*. However, this requires the collection of data from a larger group of speakers than the four talkers used in this work.

Most of the analyzed features incorporate knowledge about the human auditory system to some extent, although this does not represent a full perceptual model. We assume that a combination of the proposed perceptual training with models applied in other approaches for microscopic phoneme confusions [1, 2] could result in a further improvement of prediction quality which will be subject of future work.

5. Acknowledgements

Supported by the DFG (SFB/TRR 31 The active auditory system; URL: http://www.uni-oldenburg.de/sfbtr31). We would like to thank Thomas Brand, Tim Jürgens and Jörg-Hendrik Bach for their support and contribution to this work.

6. References

[1] Jürgens, T. and Brand, T. (2009), "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," Journal of the Acoustical Society of America, 126, pp. 2635-2648.

[2] Cooke, M. (2005), "A glimpsing model of speech perception in noise," The Journal of the Acoustical Society of America, 119, pp. 1562-1573.

[3] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (2005), "Oldenburg Logatome Speech Corpus (OLLO) for speech recognition experiments with humans and machines," in Proc. Interspeech, pp. 1273-1276.

[4] Dreschler, W. A., H, V., Ludvigson, C., and Westermann, S. (2001), "ICRA Noises: Artificial noise signals with speech-like spectral and temporal properties," Audiology, 40 (3), pp. 148-157.

[5] 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

[6] Hermansky, H. (1990), "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, Vol. 87 (4) pp. 1738-1752

[7] Hermansky, H. and Morgan, N. (1994), "RASTA processing of speech," IEEE Transactions on Speech and Audio Processing, 2, pp. 578-589.

[8] Ellis, D. (2003), "Rasta PLP in Matlab," URL: http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat.

[9] Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1995), "The HTK book," Cambridge University.

[10] Meyer, B. and Kollmeier, B. (2008), "Optimization and evaluation of Gabor feature sets for ASR," in Proc. Interspeech, pp. 906-909.

[11] Meyer, B., Wächter, M., Brand, T., and Kollmeier, B. (2007), "Phoneme confusions in human and automatic speech recognition," in Proc. Interspeech, pp. 1485-1488.

[12] Cooke, M. and Scharenborg, O. (2008), "The Interspeech 2008 Consonant Challenge," in Proc. Interspeech, pp. 1781-1784.