# Phoneme Confusions in Human and Automatic Speech Recognition

*Bernd T. Meyer, Matthias Wächter, Thomas Brand, Birger Kollmeier*

Medical Physics Section, University of Oldenburg, Germany

bernd.meyer@uni-oldenburg.de, thomas.brand@uni-oldenburg.de

## Abstract

A comparison between automatic speech recognition (ASR) and human speech recognition (HSR) is performed as prerequisite for identifying sources of errors and improving feature extraction in ASR. HSR and ASR experiments are carried out with the same logatome database which consists of nonsense syllables. Two different kinds of signals are presented to human listeners: First, noisy speech samples are converted to Mel-frequency cepstral coefficients which are resynthesized to speech, with information about voicing and fundamental frequency being discarded. Second, the original signals with added noise are presented, which is used to evaluate the loss of information caused by the process of resynthesis. The analysis also covers the degradation of ASR caused by dialect or accent and shows that different error patterns emerge for ASR and HSR. The information loss induced by the calculation of ASR features has the same effect as a deteriation of the SNR by 10 dB.
**Index Terms**: human speech recognition, automatic speech recognition, dialect, accent, phoneme confusions, MFCC

## 1. Introduction

Automatic speech recognition (ASR) has seen many advances in the last years, but the large gap between recognition of spoken language by humans and machines still prevents it from everyday use. There are several causes for the inferior performance of ASR compared to human speech recognition (HSR): Human language models are for example more sophisticated than ASR language models, as increasing ASR error rates are observed for more complex tasks [1]. Furthermore, additive or convolutive noise as well as speech intrinsic variabilities (such as, e.g., dialect or speaking rate and effort) can severely degrade ASR performance.

The aim of this study is to perform a fair comparison of human and machine phoneme recognition. For similar experimental conditions, the same speech database with non-sense syllables was used for ASR and HSR tests. Hence, human listeners were not able to exploit context knowledge and language models in ASR could be disregarded. This helps to decouple the influence of two major sources of errors in ASR, namely the front-end and the back-end. Thus, the focus is laid on the importance of phoneme classification and feature extraction. The database covers several dialects [2] which were included in the comparison. Different error patterns of the confusions of phonemes should help to identify sources of errors and to improve ASR feature extraction.

It was also investigated whether the information contained in ASR features is sufficient for human listeners to recognize speech. Therefore, the most common features in ASR (Mel-frequency cepstral coefficients / MFCCs) have been resynthesized to audible signals [3] which were presented to human test subjects. During the extensive HSR experiments, the original signals were also presented as a reference condition. The results are analyzed on a microscopic scale with phoneme confusion matrices, which have sucessfully been utilized for man-machine-comparison earlier [4, 5].

## 2. Method

### 2.1. Speech database

The speech database used for HSR and ASR experiments is the Oldenburg Logatome Corpus (OLLO) [2] which is specifically targeted at a direct comparison of speech recognition performance in HSR and ASR. It contains 150 different nonsense utterances (logatomes) spoken by 40 German and 10 French speakers. Each logatome consists of a combination of consonant-vowel-consonant (CVC) or vowel-consonant-vowel (VCV) with the outer phonemes being identical.

To provide an insight into the influence of speech intrinsic variabilites on speech recognition, OLLO covers several variabilities such as speaking rate and effort, dialect, accent and speaking style (statement and question). The dialects contained in OLLO are East Frisian, East Phalian, Bavarian and standard German. The OLLO corpus is freely available at http://sirius.physik.uni-oldenburg.de. If recognition rates of the middle phonemes are analyzed (as done here), the number of response alternatives is reduced dramatically in comparison to an open test. This makes HSR experiments tractable, since a presentation of randomized logatomes with the same outer phonemes is possible. The approach results in a closed test setup and avoids out-of-vocabulary errors. The middle phonemes of logatomes are either vowels or consonant phonemes which are listed below (represented with the International Phonetic Alphabet (IPA)).

- Consonant phonemes: /p/, /t/, /k/, /b/, /d/, /g/, /s/, /f/, /v/, /n/, /m/, /ʃ/, /ʦ/, /l/
- Vowel phonemes: /a/, /aː/, /ɛ/, /e/, /ɪ/, /i/, /ɔ/, /o/, /ʊ/, /u/

For HSR experiments, the measurement time has to be limited to a reasonable amount, which requires a subset of OLLO to be selected for speech recognition tests. A representative speaker set for which gender and dialect are uniformly distributed and which exhibits average ASR performance was chosen for the measurements. One male and one female speaker were selected from each dialect region, which results in a total of ten speakers or 1,500 utterances for the test set.

### 2.2. Experimental conditions

Speech intelligibility tests with human listeners included two conditions:

*1. Presentation of resynthesized signals:* For a fair comparison, it is investigated if the most common features in ASR contain all the information needed for humans to understand speech

on phoneme level. MFCC features are therefore resynthesized, i.e. feature vectors used internally by the speech recognizer are decoded to acoustic speech tokens. Since the calculation of MFCCs results in a loss of information, these signals sound unnatural (like synthesized speech). For example, the speaker's identity or even gender are usually not recognizable. Nevertheless, the resynthesized logatomes are perfectly understandable in the absence of noise. To allow for a valid comparison, the presented recognition scores were obtained with noisy speech. By adding noise, redundant information in the speech signal is masked, so that intelligibility is potentially decreased in contrast to an unprocessed signal. The reduction of redundancy might be particularly critical in the presence of speech intrinsic variabilities as, for example, regional dialect.

*2. Presentation of original signals*: Unaltered speech signals from the OLLO database are used as reference. A comparison with the first condition should reveal if error patterns differ and if speech information crucial for recognition is disregarded when MFCC features are calculated.

### 2.3. Calculation of MFCCs

MFCCs are a compact representation of speech signals and have been successfully applied to the problem of ASR. However, this compact representation comes at the cost of information loss: During the calculation, phase information and fine structure of the spectrum are disregarded. This is useful in the absence of noise, but may be detrimental in noisy conditions, because redundant information exploited by humans is removed. Using the phase information has, e.g., been found to be beneficial in ASR [6]. In order to calculate MFCC features from speech, signals with 16 kHz sampling frequency are windowed with 30ms Hanning windows and a frame shift of 10ms. Each frame undergoes the same processing steps: Calculation of the amplitude spectrum, reduction of the frequency resolution using a Mel-scaled filterbank and calculating the logarithm and the inverse discrete cosine transformation (IDCT) of its output. Twelve of the lower coefficients plus an additional energy feature are selected for the ASR experiments and HSR tests with resynthesized speech.

### 2.4. Re-decoding of MFCCs to speech

In order to decode these features to an acoustic speech signal, a linear neural network trained with the OLLO training set (c.f. Section 2.6) is used to construct the spectral envelope from the cepstral coefficients. Additional information such as voicing or fundamental frequency $f_g$ is not used for the calculation, since this would give human listeners an unfair advantage over ASR. Hence, an artificial excitation signal has to be used. Pilot experiments showed that intelligibility is highest when a pulse train with $f_g = 130\ Hz$ is used as excitation signal (instead of noise or a mixed noise-pulse signal). In a final step, the spectral envelope and the artificial excitation signal are combined. Due to the fixed fundamental frequency, resynthesized speech sounds artificial and tinny, but remains understandable when no noise is present. This algorithm was kindly supplied by the Katholieke Universiteit Leuven [3].

HSR scores are usually very close to 100% for the clean condition, both for the unaltered signals and the signals derived from cepstral coefficients. In [5], the lowest recognition rate observed for non-dialect speech was 99.1 percent for a similar task. This clearly demonstrates the excellence of the human auditory system, but does not allow for a valid analysis of phoneme confusions, because differences at very low or high error rates often are outside the range of reliably observable differences (ceiling effect). Hence, speech-shaped noise is used to increase the difficulty of the listening task. In case of resynthesized speech, noise is added *before* MFCCs are calculated from the original signals. Pilot measurements with one test subject showed that a ceiling effect is always observed when the same SNR is used for resynthesized and original signals, i.e. the recognition rates are either too low for the first or too high for the second condition to obtain valid and comparable results in reasonable measurement time. Based on these first measurements, the SNR for each condition was chosen to produce approximately the same recognition rates. Resynthesized and original signals were presented at an SNR of 0 dB and -10 dB, respectively.

### 2.5. Human speech recognition test setup

Five normal-hearing listeners (two male, three female) without a noticeable regional dialect participated in the HSR tests. Signals were presented in a soundproof booth via audiological headphones (Sennheiser HDA200). An online freefield equalization and randomization of logatomes was performed by the measurement software MessOL. Feedback or the possibility to replay the logatome was not given during the test procedure. In order to avoid errors due to inattentiveness, listeners were encouraged to take regular breaks. After a training phase, subjects were presented a sequence of logatomes at a level of 70 dB SPL. For each presentation, the logatome had to be selected from a list of CVCs or VCVs with the same outer phoneme and different middle phonemes. A touch screen and a computer mouse were used as input devices. In order to avoid speaker adaptation, all resynthesized signals were presented before the subjects listened to the unprocessed speech files. The HSR measurements include 1,500 presentations per listener and test condition (original and resynthesized signals), which resulted in a total of 2 x 7,500 presentations. The cumulative measurement time was about 34 hours, including pauses and instructions for listeners.

### 2.6. Automatic speech recognition test setup

ASR experiments were carried out with a Hidden Markov Model (HMM) with three states and eight Gaussian mixtures per HMM state. The system was set up to closely resemble the closed test which was used for human intelligibility tests, i.e. confusions could only occur for the middle phonemes. This was achieved by training and testing several HMM systems with each corresponding to a different outer phoneme. Additional delta and acceleration features were added to the 13 cepstral coefficients, yielding a 39-dimensional feature vector per time step. The ASR test set contained the same utterances as used in HSR experiments (ten speakers with 150 utterances each) with the exception that all repetitions of logatomes were used instead of just one (c.f. Section 2.1). Speech files from the remaining 40 speakers in OLLO were chosen for the training process, which results in a speaker independent ASR system. The frequency of phonemes and gender were equally distributed in the training and test set. ASR recognition scores were obtained for different SNRs, for which a speech-shaped noise was added to the utterances; the same SNR was used for training and test, resulting in a matched training-test-condition.

## 3. Results

Overall HSR and ASR results for several conditions are presented in Table 1: Due to the adjustment of the SNR (c.f. Sec-

tion 2.4) the total recognition scores for both HSR conditions are very similar (shaded elements). Consonants are recognized slightly better than vowels in case of resynthesized signals, but intelligibility of consonants is lower for unprocessed signals, whereas the scores for vowels are *higher* in spite of the lower SNR (color-inverted elements). Regarding dialect, no large differences between the conditions can be observed: Although the order of 'no dialect' and 'East Frisian' is swapped for both HSR conditions, differences in recognition scores are much smaller than for vowel and consonant recognition. The performance for the dialects decreases in the order Bavarian, East Phalian and French.

Confusion matrices (CMs) characterize how often a presented phoneme was recognized or confused with response alternatives (see Figs. 1 to 3). The matrices are based on the complete measurements for the corresponding condition, i.e., CM scores are averaged over all dialects. For the HSR measurements, all presentations in a row correspond to 250 or 400 single presentations of consonants and vowels, respectively. In case of ASR, this corresponds to 150 (consonant recognition) or 240 (vowel recognition) utterances.

| | | HSR | | ASR | | | |
|---|---|---|---|---|---|---|---|
| | | orig -10 dB | resynth 0 dB | clean | 10 dB | 0 dB | -10 dB |
| Total | | 74.0 | 73.8 | 82.1 | 79.3 | 68.5 | 34.0 |
| Consonants | | 65.7 | 74.3 | 85.1 | 80.2 | 59.5 | 21.6 |
| Vowels | | 81.3 | 73.3 | 79.4 | 78.5 | 76.4 | 44.9 |
| Dialect / Accent | No dialect | 81.5 | 77.5 | 88.4 | 87.0 | 75.4 | 42.6 |
| | East Frisian | 80.9 | 79.2 | 84.5 | 82.5 | 72.4 | 34.0 |
| | Bavarian | 77.6 | 75.1 | 79.1 | 75.4 | 66.8 | 36.2 |
| | East Phalian | 70.2 | 71.3 | 84.1 | 78.5 | 68.0 | 30.8 |
| | French | 59.7 | 65.7 | 74.2 | 73.2 | 59.8 | 26.3 |

Table 1: Recognition rates in % for HSR (original signals, presented at -10 dB SNR and resynthesized signals, presented at 0 dB SNR) and ASR (at different SNRs). In the first three rows, the average accuracy ('total') as well as consonant (VCV) and vowel (CVC) rates are reported. The subsequent rows show the recognition performance depending on dialect and accent.

## 4. Discussion

A direct comparison of ASR and HSR performance shows that human speech recognition is superior to the ASR system under equal conditions, as presented in Table 1. The total HSR and ASR accuracies at an SNR of -10 dB are 74.0% and 34.0%, respectively (encircled elements), which corresponds to a relative increase of the word error rate (WER) of 154%. The gap narrows if the information for human listeners is limited to the information content of MFCCs: For resynthesized signals at 0 dB SNR, the recognition score is 73.8% and the corresponding ASR accuracy is 68.5%, resulting in an WER increase of 20.0%.

The SNRs for both HSR conditions were chosen so that average recognition rates are similar. The choice of SNRs was based on the presentation of only few test lists to one human listener and proved to be reasonable for other test subjects as well, as the overall accuracies are very close to each other: The average scores were 73.8% (resynthesized signals) and 74.0% (original signals). Therefore, the information loss caused by MFCCs can be expressed in terms of the signal-to-noise ratio, i.e. the SNR of resynthesized signals has to be 10 dB higher in order to obtain similar recognition performance. This is consistent with the model of human speech perception using an auditory model as front-end to ASR as presented in [8].

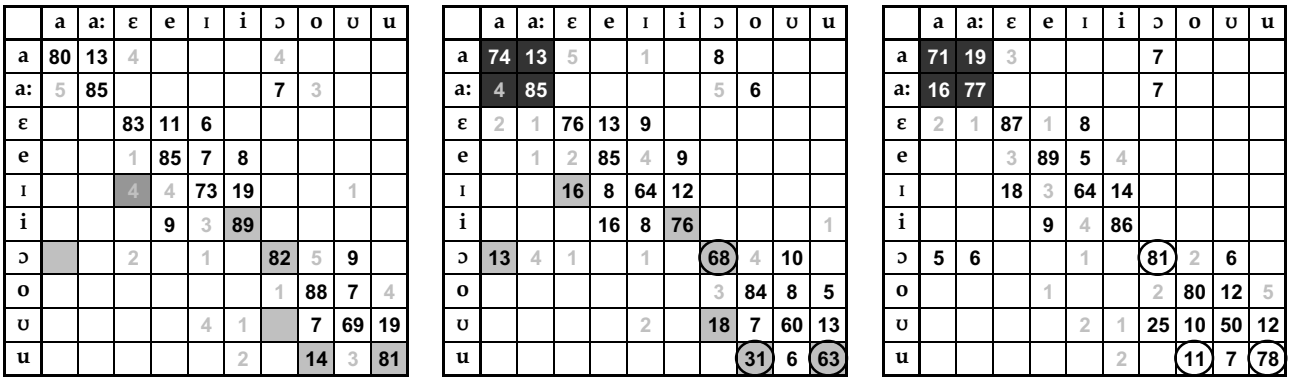
The vowel recognition rates for both HSR conditions show that the information loss during feature calculation is particularly problematic for vowels, as accuracy drops by 10% when using resynthesized instead of original signals even though the SNR is *lower* for the original signals. Although MFCCs have been found to encode the spectral shape of vowels well, the reduced frequency resolution may result in inferior differentiation between proximate formants compared to human listeners.

| | p | t | k | b | d | g | s | f | v | n | m | ʃ | ts | l |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 58 | 4 | 9 | 10 | 1 | 5 | | 4 | 8 | | | | | |
| t | 2 | 74 | 5 | | 10 | 4 | | | | | | | 4 | |
| k | 4 | 2 | 70 | | 1 | 18 | | | 3 | | | | 1 | |
| b | 11 | | 2 | 44 | 3 | 13 | | | 26 | | 1 | | | 2 |
| d | | 2 | | 4 | 74 | 13 | | | 3 | 3 | | | | |
| g | 3 | | 11 | 2 | 2 | 73 | | | 5 | 2 | 1 | | | 1 |
| s | | | | | | | 87 | 5 | | | | 4 | 1 | 2 |
| f | 1 | | 1 | | | | 1 | 89 | 7 | | | 1 | | |
| v | 1 | | | 16 | 1 | 6 | | 3 | 70 | | 2 | | | 1 |
| n | | | | 1 | 4 | 1 | | | 2 | 61 | 12 | | | 19 |
| m | | | | 4 | | 3 | | | 5 | 12 | 58 | | | 18 |
| ʃ | | | | | | | 1 | 2 | | | | 98 | | |
| ts | | 3 | | | | | 2 | | | | | | 95 | |
| l | | | | 2 | 4 | 2 | | | 2 | 4 | | | | 88 |

Figure 1: Confusion matrix (CM) for consonant phonemes, derived from human speech recognition tests with resynthesized speech at an SNR of 0 dB. The matrix element $C_{ij}$ denotes how often the phoneme in row $i$ was classified as the phoneme in column $j$. Rows are normalized to 100%. Matrix elements with a value of zero are not plotted and elements $< 5$ are plotted in light gray for reasons of readability. Inverted elements denote large differences between this CM and Fig. 2.

| | p | t | k | b | d | g | s | f | v | n | m | ʃ | ts | l |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 54 | 1 | 11 | 16 | 7 | 3 | | 1 | 5 | 1 | | | 1 | |
| t | 9 | 27 | 9 | 3 | 17 | 1 | | | 1 | | | | 33 | |
| k | 16 | 2 | 55 | 3 | 7 | 9 | | 1 | 1 | | | | 6 | |
| b | 15 | 1 | 3 | 39 | 8 | 7 | | 2 | 16 | 5 | 3 | | 1 | |
| d | 3 | 3 | 1 | 6 | 60 | 9 | | | 4 | 5 | 2 | | 3 | 5 |
| g | 9 | 1 | 14 | 9 | 13 | 39 | | | 8 | 4 | 2 | | | 2 |
| s | 1 | 1 | 1 | | 1 | | 83 | 5 | | 1 | 1 | 3 | 5 | |
| f | 7 | | 1 | 7 | 1 | | 1 | 76 | 6 | | | 1 | 1 | |
| v | 3 | 1 | 2 | 27 | 7 | 8 | | 9 | 36 | 2 | 3 | | | 3 |
| n | | 1 | 1 | 3 | | | 1 | | 3 | 61 | 18 | | 1 | 11 |
| m | 1 | | 1 | 5 | 1 | 1 | | | 4 | 31 | 54 | | | 2 |
| ʃ | | | | | | | 1 | 3 | | | | 96 | | |
| ts | | 9 | 1 | | | | 1 | 1 | | | 1 | | 89 | |
| l | 1 | | 1 | 3 | 2 | 2 | | | 2 | 19 | 3 | 1 | | 66 |

Figure 2: CM for consonant phonemes, derived from ASR experiments for which training and test data at 0 dB SNR were used. See Fig. 1 for details.



| | a | a: | ɛ | e | ɪ | i | ɔ | o | ʊ | u |
|---|---|---|---|---|---|---|---|---|---|---|
| a | 80 | 13 | 4 | | | | 4 | | | |
| a: | 5 | 85 | | | | | 7 | 3 | | |
| ɛ | | | 83 | 11 | 6 | | | | | |
| e | | | 1 | 85 | 7 | 8 | | | | |
| ɪ | | | 4 | 4 | 73 | 19 | | | 1 | |
| i | | | | 9 | 3 | 89 | | | | |
| ɔ | | | 2 | | 1 | | 82 | 5 | 9 | |
| o | | | | | | | 1 | 88 | 7 | 4 |
| ʊ | | | | | 4 | 1 | | 7 | 69 | 19 |
| u | | | | | | 2 | | 14 | 3 | 81 |

| | a | a: | ɛ | e | ɪ | i | ɔ | o | ʊ | u |
|---|---|---|---|---|---|---|---|---|---|---|
| a | 74 | 13 | 5 | | 1 | | 8 | | | |
| a: | 4 | 85 | | | | | 5 | 6 | | |
| ɛ | 2 | 1 | 76 | 13 | 9 | | | | | |
| e | | 1 | 2 | 85 | 4 | 9 | | | | |
| ɪ | | | 16 | 8 | 64 | 12 | | | | |
| i | | | | 16 | 8 | 76 | | | | 1 |
| ɔ | 13 | 4 | 1 | | 1 | | 68 | 4 | 10 | |
| o | | | | | | | 3 | 84 | 8 | 5 |
| ʊ | | | | | 2 | | 18 | 7 | 60 | 13 |
| u | | | | | | | | 31 | 6 | 63 |

| | a | a: | ɛ | e | ɪ | i | ɔ | o | ʊ | u |
|---|---|---|---|---|---|---|---|---|---|---|
| a | 71 | 19 | 3 | | | | 7 | | | |
| a: | 16 | 77 | | | | | 7 | | | |
| ɛ | 2 | 1 | 87 | 1 | 8 | | | | | |
| e | | | 3 | 89 | 5 | 4 | | | | |
| ɪ | | | 18 | 3 | 64 | 14 | | | | |
| i | | | | 9 | 4 | 86 | | | | |
| ɔ | 5 | 6 | | | 1 | | 81 | 2 | 6 | |
| o | | | | 1 | | | 2 | 80 | 12 | 5 |
| ʊ | | | | | 2 | 1 | 25 | 10 | 50 | 12 |
| u | | | | | | 2 | | 11 | 7 | 78 |

Figure 3: CMs for vowel phonemes, derived from HSR tests with original signals at -10 dB SNR (left panel), from resynthesized signals at 0 dB SNR (middle panel) and from ASR experiments at 0 dB SNR (right panel). Gray-shaded elements highlight degradations that emerge when resynthesized signals instead of the original ones are used. Encircled cells show improvements of ASR compared to resynthesized features, while color-inverted elements show degradations. See Fig. 1 for details.

The performance drop may also be caused by discarding the phase component. Corresponding findings have been obtained for ASR where performance was improved by exploiting phase information [6] and in HSR when the audible information was limited to the power spectrum of noisy signals [7]. Preliminary measurements have shown that the information contained in MFCCs is sufficient to recognize speech in the absence of noise, since the intelligibility in HSR is not degraded when using resynthesized signals instead of the original ones. However, the presented measurements in noise clearly show that during the calculation of MFCCs a significant amount of useful information is removed. These conclusions are based on the assumption that the decoding algorithm for MFCCs uses all the information contained in MFCCs.

A comparison of the CMs shows that in some cases recognition performance is severely degraded when presenting resynthesized signals instead of original ones (gray-shaded elements in Fig. 3). This suggests that too much redundancy is removed by the feature calculation process so that the according phonemes cannot be distinguished. In other cases, ASR performance is higher than HSR with resynthesized features (encircled elements) which suggests that either the feature information is not completely made audible by the decoding algorithm or the auditory model of the ASR back-end covers the acoustic space optimally and is thus superior to HSR for these microscopic confusions. Finally, color-inverted elements in the CMs (Figs. 1-3) denote cases in which ASR performs worse than humans with resynthesized signals which means that the feature information is not optimally exploited by the back-end. Many of the observed errors can either be attributed to feature extraction or to the back-end, which might be helpful to improve ASR.

## 5. Conclusions

1. Even for the relatively simple task of phoneme classification, the difference between HSR and ASR remains considerably large: The increase of relative WER is larger than 150% at -10 dB SNR. If the information contained in MFCC features is resynthesized and presented to human listeners, the gap narrows, but error rates are still 20% higher for ASR.

2. The information loss caused by the calculation of Mel-frequency cepstral coefficients can be expressed in terms of the signal-to-noise ratio: Similar recognition results in HSR are obtained when the SNR is 10 dB higher for resynthesized signals instead of unaltered speech files.

3. Regarding dialect, no major differences between resynthesized and original signals can be observed. This suggests that information is equally well encoded in MFCCs for the dialects which were subject of the analysis.

4. For the analyzed SNR conditions, information needed to distinguish between vowel phonemes seems to be encoded suboptimally by MFCCs which may be caused by missing phase information or reduced spectral resolution.

*Acknowledgements*: Supported by the DFG (SFB/TR 31 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbtr31).

## 6. References

[1] Lippmann, R.P., "Speech Recognition by Machines and Humans", Speech Communication 22 (1) 1-15, 1997.

[2] Wesker T. et al., "Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines", In Proc. Interspeech 2005, Lisbon, Portugal, pp. 1273-1276.

[3] Demuynck, K., Garcia, O. and Dirk Van Compernolle., "Synthesizing Speech from Speech Recognition Parameters", In Proc. ICSLP 2004, volume II, pages 945–948.

[4] Sroka, J.J. and Braida, L.D., "Human and Machine Consonant Recognition", Speech Communication 45 (401-423), 2005.

[5] Meyer, B.T. and Wesker, T., "A Human-Machine Comparison in Speech Recognition Based on a Logatome Corpus", Workshop on Speech Recognition and Intrinsic Variation, May 2006, Toulouse, France.

[6] Schlüter, R. and Ney, H., "Using Phase Spectrum Information for Improved Speech Recognition Performance", In Proc. ICASSP, 2001.

[7] Peters, S.D., Stubley and P., Valin, J.-M., "On the Limits of Speech Recognition in Noise", In Proc. ICASSP, 1999.

[8] Jürgens, T., Brand, T. and Kollmeier, B., "Modelling the Human-Machine Gap in Speech Reception: Microscopic Speech Intelligibility Prediction for Normal-Hearing Subjects with an Auditory Model", this issue.