# A HUMAN-MACHINE COMPARISON IN SPEECH RECOGNITION BASED ON A LOGATOME CORPUS

*Bernd Meyer*[1], *Thorsten Wesker*[1], *Thomas Brand*[1], *Alfred Mertins*[2], *Birger Kollmeier*[1]

Department of Physics, [1]Medical Physics, [2]Signal Processing Group
Carl von Ossietzky University of Oldenburg, Germany
medi-ollo@listserv.uni-oldenburg.de

## ABSTRACT

In this study, a fair comparison of human and machine speech recognition is established by using the same paradigms for human speech recognition (HSR) and automatic speech recognition (ASR). In order to ensure equal conditions, a speech database specifically designed for this task is used. The results for HSR and ASR are broken down into several intrinsic variabilities like speaking rate, speaking effort and dialect. Across all conditions, ASR error rates are at least 300 % higher than those of humans, even though no contextual knowledge can be exploited. A more detailed analysis of errors in HSR and ASR is carried out by decomposing speech into its phonetic features like voicing or manner and place of articulation. Confusion matrices for these features show that voicing information is crucial to distinguish between certain consonants. The most prominent features for ASR often neglect voicing information, which might contribute to the large gap in performance between HSR and ASR.

## 1. INTRODUCTION

While ASR has seen many advances in recent years, the error rates of machines are still an order of magnitude higher than those of humans. This leads the interest to the sophisticated mechanisms that give the human sense of hearing its excellence in comprehension and error correction. To learn more about these mechanisms, it is useful to have a direct comparison between the recognition abilities of both HSR and ASR. Former studies [2] support these arguments, but most of the comparisons fail to establish equal conditions (e.g. because different corpora were used for ASR and HSR). We eliminated most of the unequal conditions by compiling a context free speech database, on which both sides could perform unbiased tests and which satisfies requirements for ASR and HSR tests. This logatome speech database [1] is briefly described in the first paragraph. The experimental setup, designed to carry out an unbiased comparison of performance, is specified in Section 3. Results for HSR and ASR are presented and an analysis based on decomposing speech into phonetic features is described in detail (Section 5). Future experiments that will use the same data analysis scheme and which are based on speech resynthesized from ASR features are presented in Section 6. Finally, we draw the conclusions from the comparison experiments in Section 7.

## 2. SUMMARY LOGATOME SPEECH DATABASE

The OLdenburg LOgatome Corpus (OLLO)[1] is specially designed to allow for an unbiased human-machine comparison by comparing the recognition capabilities for individual phonemes that are embedded in logatomes, namely three-phoneme sequences with no semantic information [2]. These context-free phoneme-sequences were combined either as CVC (consonant-vowel-consonant) or VCV (vowel-consonant-vowel) logatomes. A balanced set of target-phonemes important for human and automatic speech recognition has been chosen, drawing on pilot ASR studies and cross-fertilization from the field of human speech intelligibility testing. The preliminary studies resulted in 70 VCVs and 80 CVCs, for example:

- VCV: ataː, afaː, adaː     ɛtə, ɛfə, ɛdə
- CVC: tat, tʊt, tɛt, taːt, tut, tet
  faf, fʊf, fɛf, faːf, fuf, fef

The second purpose of the corpus is to cover speech intrinsic variabilities that affect recognition rates of humans and machines. OLLO contains items with six different articulation characteristics. These characteristics are:

- speaking rate (slow, normal, fast)
- speaking effort (quiet, normal, loud)
- speaking style (statement or question)

Dialect is another speech intrinsic variability of the corpus, as OLLO contains recordings of 40 speakers (20 male and 20 female) from four different German regions with distinct dialects. These dialects are East Frisian (EF), East Phalian (EP), Bavarian (BV) and standard German.

Additionally, sets of 72 words and 20 sentences were recorded per speaker, both phonetically balanced. Those are designed for ASR training purposes and speaker adaptation. All speakers were advised to speak in a natural manner and not to suppress their dialect. To avoid systematic errors the variabilities were recorded in random order. Each speaker recorded 150 logatomes in six variabilities and three repetitions. This resulted in 2700 utterances per speaker and a whole of 107,000 logatomes or 43.3 hours of speech in OLLO.

The recording of the raw data was carried out with professional digital studio hardware (condenser-microphone AKG C1000 S, RME QuadMic microphone pre-amplifier, RME Hammerfall AD-Converter at 44,1 kHz and 32 bit resolution) in unechoic recording chambers. A specially developed software tool controlled the speaker instructions, the recording and storage of the speech data. Finally, the following signal processing steps were applied:

- limitation of silence at beginning and end of signal to 500 ms
- normalization of amplitudes to 99 %
- storage in 16 bit resolution
- low-pass filtering with 8 kHz cutoff frequency
- downsampling to 16 kHz

## 3. EXPERIMENTAL SETUP

### 3.1. Paradigms for ASR and HSR

The OLLO database contains a large amount of test items that only a machine can process in acceptable time. For the recognition experiments with humans, it is necessary to reduce the test set and the number of response alternatives. This leads to the necessity to change the test paradigms from semi-open to closed tests.

In the semi-open test, 150 response alternatives (70 VCVs + 80 CVCs) exist for each logatome. When the results are plotted in a matrix of confusion, all elements can be mixed up and errors may appear anywhere beside the diagonal. Certain errors (like confusions between short and long vowels) can be identified easily if the matrix is ordered in a systematic way.

For speech intelligibility tests, it is not tractable to let the listener choose between 150 alternatives when a large amount of items is to be tested. Therefore, a forced choice test is carried out, where the intelligibility of the middle phoneme is tested so that the number of response alternatives is reduced. For example, listeners have to choose between logatomes like ”ʊdu”, ”ʊtu”, ”ʊku”, ”ʊgu”, (with 14 alternatives for VCVs and 10 for CVCs). The outer phonemes are the same for all test items of the reduced set of answers.

In order to have similar paradigms in HSR and ASR, closed test lists were used for ASR experiments as well. This was realized by using 14 different HMM systems, corresponding to the 14 outer phonemes (5 outer vowels + 9 outer consonants). Each system was trained and tested with speech files with same outer phonemes, which restricts ASR errors to misclassification of the middle phoneme. This reduces response alternatives and therefore improves recognition scores.

### 3.2. OLLO subset for HSR experiments

It is also not tractable for humans to do recognition experiments on the whole corpus with over 100,000 logatomes. Due to this fact, a representative subset, covering variabilities and dialects, had to be found. For the HSR experiments, utterances of one male speaker were selected for each dialect region. The choice was based on

- speaking quality (best realization of the logatomes)
- recording mistakes / missing files
- strength of dialect

Logatomes without any dialect are presented in all variabilities to the listeners. Logatomes uttered by the three chosen dialect speakers are only presented in condition ’statement’.

The subset size is a tradeoff between having better statistics and a reasonable time consumption for the listener. There are more outer phonemes in the CVC sets than in the VCV sets. To get similar statistics in the confusion matrices for the middle phonemes, the HSR test lists include one repetition of the CVC set and two repetitions of the VCV set, i.e. 5 long and 5 short vowels with 8 outer phonemes result in 80 CVCs. 14 consonants with 5 outer vowels result in 70 VCVs. Therefore, the subset consists of 140 VCVs and 80 CVCs for each variability and dialect.

### 3.3. Experimental Setup for ASR

ASR experiments were carried out using a Hidden Markov Model (HMM), where a three-state-model was used for each phoneme and a word model exists for each logatome. The HMM was built with the hidden Markov toolkit (HTK), available at http://htk.eng.cam.ac.uk/. It was trained and tested with logatomes from the OLLO database. As defined in the OLLO corpus segmentation files, 5 speakers from each dialect region are contained in the training and test sets. The sets have the same number of utterances and exhibit the same statistics regarding gender and intrinsic variabilities (including dialect, speaking rate and speaking effort). Since speakers are either in the training or in the test set, the task is to perform speaker-independent recognition. A number of features have been tested; the results presented here were obtained with mel frequency cepstral coefficients (13 MFCCs with 25 ms window length, 10 ms frame shift + $\Delta$ + $\Delta\Delta$-features), which produced lowest error rates.

## 4. EXPERIMENTAL RESULTS

The overall recognition rates in the ASR experiments with closed test lists was 74.0 %. Results are plotted with respect to the different variabilities contained in OLLO and compared to human recognition rates in Figs. 1 and 2.

For HSR experiments, signals were presented with a pair of Sennheiser HDA 200 headphones at a level of 70 dB SPL in

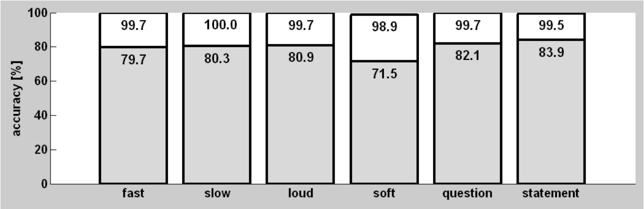

**Fig. 1**. Recognition rates for ASR (grey shading) and HSR (no shading) depending on speech intrinsic variabilities.

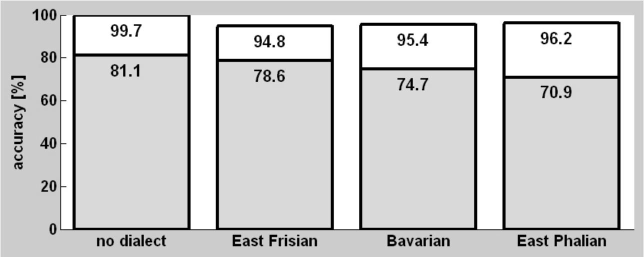

**Fig. 2**. Recognition rates for ASR (grey shading) and HSR (no shading) depending on regional dialect.

a soundproof booth. Six normal hearing listeners participated in the intelligibility tests. Recognition rates are very high for the clean speech condition. The averaged recognition rates are best for standard German and slightly decrease for the dialects (see Fig. 3).

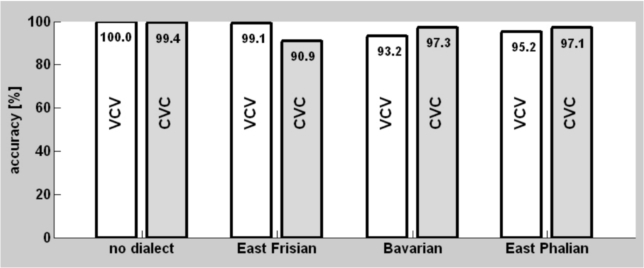

**Fig. 3**. HSR performance for consonant and vowel recognition (CVC and VCV, respectively) in clean speech, depending on dialect.

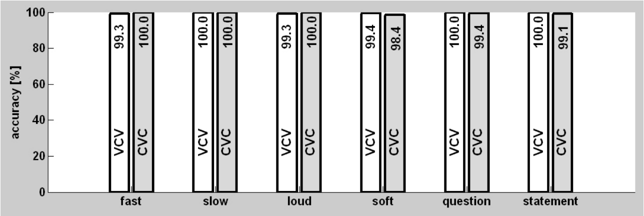

**Fig. 4**. HSR performance for consonant and vowel recognition (CVC and VCV, respectively) in clean speech, depending on variability.

In Fig. 4, the recognition results for clean speech are sorted depending on variabilities. The accuracies for all conditions are nearly 100 %, a small variation about 1 % is noticeable. However, these differences are not statistically significant.

Obviously, there are too few errors in clean speech experiments with humans to gain much insight into which tasks are more and which are less difficult for humans. Since we try to learn from the mechanisms in human speech perception we need to complicate the conditions for humans for finding the critical points in their information transmission. Due to this fact we conducted experiments with added noise at a level of 70 dB SPL and an SNR of 0 dB. The results are presented in Figs. 5 and 6. Results for CVCs and VCVs are plotted separately and can be compared with the results for clean speech in Figs. 3 and 4. In this paradigm, it is much more difficult for humans to recognize consonants, as one can see from the decrease of recognition rates for VCVs in noisy condition. The recognition rates for CVCs stay on the same level as in clean speech. These results suggest to carry out an analysis of the errors for VCVs in a more detailed way, which is described in Section 5.

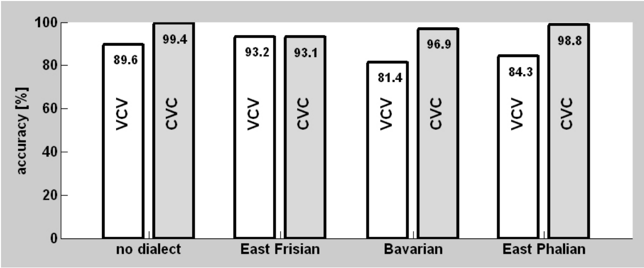

**Fig. 5**. HSR performance for consonant and vowel recognition (CVC and VCV, respectively) in noisy speech, depending on dialect.

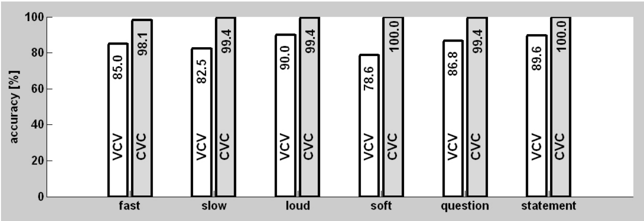

**Fig. 6**. HSR performance for consonant and vowel recognition (CVC and VCV, respectively) in noisy speech, depending on variability.

## 5. ANALYSIS OF RESULTS & DISCUSSION

### 5.1. General Results

Across all conditions, humans outperform ASR systems. The differences are clearly visible in the noise-free conditions, where the average human performance is 99.6 % and the ASR performance is 74.0 %. For the East Frisian dialect, the difference in terms of absolute error rates is only 16.2 %. Still, even in this condition, the relative ASR error rates are more than 300 % higher than the HSR error rates.

A comparison of HSR and ASR performance depending on dialects (see Fig. 2) shows that for both experiments standard German is the condition with best performance (HSR: 99.7 % accuracy; ASR: 81.1 % accuracy). The ASR accuracy is reduced by 2.5 to 10.2 % absolute for the dialects. For the HSR test, an average performance drop of 4 % is observed. Other than for the ASR experiment, the differences between dialects are not significant.

The ASR results depending on variabilities (Fig. 1) show that logatomes spoken as statements lead to best results with almost 84 % accuracy. The condition ”question” decreases accuracy by only 1.8 %. For the variabilities ”loud”, ”slow” and ”fast” another 2 % drop is observed. The conditions ”soft” is the most problematic for the HTK system, since the absolute accuracy is more than 10 % worse than for the ”statement” condition. It is difficult to compare the corresponding HSR results with this performance, because the differences between intrinsic variabilities are not significant, as stated earlier.

### 5.2. Consonant Confusion and Phonetic Features

Most of the errors in HSR tests occurred when VCVs were presented, i.e. when a consonant had to be classified. This becomes most obvious in the HSR noisy test condition (c.f. Section 4). Therefore, a detailed analysis of consonant confusions is carried out. The confusion matrices are shown in Figs. 7 and 8.

| | p | t | k | b | d | g | s | f | v | n | m |
|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 97.1 | 0 | 0 | 2.5 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 |
| t | 0 | 99.5 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| k | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| b | 2 | 0 | 0 | 94.6 | 0 | 0 | 0 | 0 | 3.4 | 0 | 0 |
| d | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| g | 0 | 0 | 2.4 | 0 | 0 | 97.6 | 0 | 0 | 0 | 0 | 0 |
| s | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| f | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 98 | 0 | 0 | 0 |
| v | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| n | 0 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 99.5 | 0 |
| m | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

**Fig. 7**. Matrix of confusion for consonants, obtained from human listening tests (accuracy 98.7 %)

| | p | t | k | b | d | g | s | f | v | n | m |
|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 88.7 | 1.6 | 2 | 4.4 | 1.4 | 1.4 | 0 | 0 | 0.3 | 0.1 | 0.1 |
| t | 4.8 | 84.2 | 3.7 | 0.1 | 5.9 | 1.2 | 0.1 | 0 | 0.1 | 0 | 0 |
| k | 1.8 | 1.7 | 91 | 0.1 | 0.3 | 5 | 0 | 0 | 0.1 | 0 | 0 |
| b | 21.3 | 0.5 | 0.9 | 59.2 | 6.8 | 4.7 | 0.1 | 0.2 | 4.6 | 0.2 | 1.5 |
| d | 1.9 | 11.2 | 1.9 | 1.6 | 71.6 | 9.9 | 0.2 | 0 | 0.6 | 1 | 0.1 |
| g | 0.9 | 0.6 | 24.3 | 1.3 | 5.3 | 64.9 | 0.1 | 0.1 | 0.9 | 1.1 | 0.5 |
| s | 0.2 | 0.5 | 0 | 0.2 | 1.3 | 0.7 | 91.3 | 1.6 | 4 | 0.2 | 0.1 |
| f | 0.5 | 0.6 | 0.2 | 0.2 | 0.3 | 0.2 | 3.9 | 92.8 | 1.1 | 0.1 | 0 |
| v | 2.4 | 0.1 | 1 | 4.6 | 2.1 | 1.9 | 0.3 | 7.7 | 76.6 | 0.6 | 2.6 |
| n | 0 | 0.1 | 0.4 | 0 | 1.4 | 1.2 | 0.1 | 0.1 | 0.8 | 89 | 7 |
| m | 0.1 | 0.1 | 0.9 | 0.3 | 0.3 | 1.1 | 0.3 | 0.3 | 2.1 | 11.7 | 82.8 |

**Fig. 8**. Matrix of confusion for consonants, obtained from ASR experiments (accuracy 81.1 %)

Our method of data analysis is based on the work by Christiansen and Greenberg [3]: The acoustic cues important for consonant identification are analyzed by decomposing consonants into their phonetic features. Eleven consonants are partitioned into three (overlapping) groups on the basis of the phonetic properties of voicing, articulatory manner and place of articulation, as illustrated in Table 1. Voicing refers to the presence (or absence) of glottal vibration. Manner refers to the mode of articulatory production (stop, nasal, fricative) and place of articulation refers to the locus of articulatory constriction (anterior, medial, posterior). Voicing is a binary distinction, while manner and place both have three class distinctions.

| consonant | p | t | k | b | d | g | s | f | v | n | m |
|---|---|---|---|---|---|---|---|---|---|---|---|
| voicing | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| manner | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 |
| place | 0 | 1 | 2 | 0 | 1 | 2 | 1 | 0 | 0 | 2 | 1 |

**Table 1**. Phonetic features for the 11 consonants used in the experiments. Voicing is a binary feature, while manner and place are ternary features.

Confusion matrices for each phonetic feature dimension can be derived from the original consonant confusion matrix: For each element in the original matrix, the values of the corresponding phonetic features are determined according to Table 1 (e.g. for the confusion p → b, the values would be unvoiced → voiced). The number of confusions is added to the appropriate matrix element (e.g. element (U,V) in the voicing matrix).

Consonant identification was scored in terms of how well the according phonetic features (voicing, manner and place) were decoded. If a consonant was correctly classified, its constituent phonetic features were also correctly decoded. If a consonant was incorrectly classified, not all phonetic features were necessarily also misclassified. For example, if the logatome 'apa:' was presented, but 'aka:' was classified, then the phonetic features 'voicing' and 'manner' were correctly decoded. Figs 9 and 10 show confusion matrices for the phonetic features.

| | S | F | N |
|---|---|---|---|
| S | 99.3 | 0.7 | 0 |
| F | 0 | 100 | 0 |
| N | 0.2 | 0 | 99.8 |

(a) Manner

| | A | M | P |
|---|---|---|---|
| A | 99.6 | 0.4 | 0 |
| M | 0 | 100 | 0 |
| P | 0 | 0 | 100 |

(b) Place

| | V | U |
|---|---|---|
| V | 99.3 | 0.7 |
| U | 0.6 | 99.4 |

(c) Voicing

**Fig. 9**. Matrices of confusion for different phonetic features, obtained from human listening tests. a) Manner of articulation, total accuracy: 99.7 %. Meaning of labels: S = Stop, F = Fricative, N = Nasal. b) Place of articulation, total accuracy: 99.9 %. Meaning of labels: A = Anterior, M = Medial, P = Posterior. c) Voicing, total accuracy: 99.3 %. Meaning of labels: V = Voiced, U = Unvoiced.

| | S | F | N |
|---|---|---|---|
| S | 97.9 | 1.3 | 0.8 |
| F | 5.7 | 93.1 | 1.2 |
| N | 2.9 | 1.9 | 95.3 |

(a) Manner

| | A | M | P |
|---|---|---|---|
| A | 90.9 | 6.3 | 2.9 |
| M | 5.8 | 89.6 | 4.7 |
| P | 2.8 | 4.6 | 92.6 |

(b) Place

| | V | U |
|---|---|---|
| V | 86.9 | 13.1 |
| U | 5.7 | 94.3 |

(c) Voicing

**Fig. 10**. Matrices of confusion for different phonetic features, obtained from ASR experiments. a) Manner of articulation, total accuracy: 95.4 %. b) Place of articulation, total accuracy: 91.0 %. c) Voicing, total accuracy: 90.6 %. Meaning of labels as in Fig. 9.

### 5.3. Information Transmission

In order to compute the amount of information transmission associated with a particular feature and stimulus condition it is necessary to measure the relationship between a specific stimulus $x$ and the response categories $y$ without the influence of a response bias. To explain the method, let us assume that the input variable $x$ can assume the discrete values $i = 1, 2, , k$ (e.g. the index number of a spoken logatome or value of a phonetic feature) with probabilities $p_i$. The response variable $y$ can assume the values $j = 1, 2, , m$ (e.g. the index of a recognized logatome or the value of a phonetic feature) with probabilities $p_j$. A joint probability $p_{ij}$ is defined as the probability of the joint occurrence of an input value $i$ and the output value $j$. The information transmission can then be computed using the expression

$$T(x, y) = -\sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}}$$

as described in [5]. This method can be used to obtain the information transmission for each phonetic feature (Voicing, Manner and Place) by determining $T(x, y)$ from the confusion matrices for the phonetic features. The results of the analysis are presented in Table 2.

ASR and HSR results based on tests with closed lists and clean speech have been used for the comparison. As for the matrix of confusion for logatomes, spoken features correspond to rows and recognized features to columns. For reasons of comparability, performance is reported in terms of relative accuracy, i.e. elements of each matrix row add up to 100 %. Shaded elements point out features or consonants with the highest relative error rates.

The matrices of confusion for consonants indicate that in ASR experiments the most frequent errors are 'g' → 'k' (i.e. spoken: 'g', recognized 'k') and 'b' → 'p', which are also problematic in human listening tests. On the other hand, the errors 'b' → 'v' and 'f' → 's' are prominent in the HSR confusion matrix, while they are not as noticeable in the corresponding ASR matrix. Errors that often occur in ASR are 'm' → 'n' and 'd' → 'g', which weren't confused in HSR tests a single time.

| | Total | Manner | Voicing | Place |
|---|---|---|---|---|
| Source Entropy H(x) | 3.46 | 1.43 | 0.99 | 1.5 |
| HSR: T(x,y) | 3.37 | 1.40 | 0.94 | 1.48 |
| ASR: T(x,y) | 2.41 | 1.17 | 0.55 | 0.98 |
| HSR: T(x,y)/H(x) | 0.97 | 0.98 | 0.94 | 0.99 |
| ASR: T(x,y)/H(x) | 0.70 | 0.81 | 0.55 | 0.66 |

**Table 2**. Amount of information transmitted in bits $T(x, y)$ and source entropy $H(x)$ in bit, calculated from the matrix of confusion for consonants (corresponding to label 'Total'), Manner, Voicing and Place. In addition, the relative throughput $T(x, y)/H(x)$ is listed.

A look at the phonetic features reveals that voicing and place of articulation exhibit the worst ASR recognition rates. The voicing and place features have a recognition score of 90.9 % and 91.0 %, respectively, while manner is classified correctly in 95.4 % of all cases. Due to the type of analysis, the scores for phonetic features are better or equal to the original recognition rate, based on the matrix of confusion for consonants. In the HSR tests, this original score was at 98.7 %, and the accuracies for phonetic features are in the range from 99.3 to 99.9 %. The largest difference of relative throughput $T(x, y)/H(x)$ between ASR and HSR can be observed for the voicing feature with a value of 0.39. The differences for the features place and manner are 0.33 and 0.17, respectively.

With such high recognition scores, it is difficult to extract statistically valid information from the data. However, from the ASR results it can be concluded that voicing information is important for good ASR performance: The voicing feature is often considered as redundant for ASR, because speech usually remains intelligible, even when voicing is discarded and noise is used as excitation signal. Our data suggests that the voicing feature is not redundant but is needed to distinguish between consonants like 'b' and 'p' or 'g' and 'k'. The fact that MFCC features neglect voicing information and the errors 'b' → 'p' and 'g' → 'k'are the most frequent in ASR experiments support this argument.

## 6. PERSPECTIVE ON FUTURE WORK

The different recognition scores for humans and machines complicate a comparison of results. With HSR results close to 100 %, a large number of logatomes has to be contained in listening tests in order to ensure results with statistical significance.

A possible solution to this problem is to use resynthesized speech in listening experiments. By reversing the signal processing of ASR systems, feature vectors used internally by the speech recognizer can be decoded to speech. Using the results of the decoding process for HSR tests could give an answer to the question whether all the information needed for recognition of speech is still present in the feature vectors. Furthermore, the process of resynthesis is likely to introduce distortions to the speech signal, so that human recognition performance is expected to drop and comparable results are obtained in ASR and HSR tests. An algorithm to calculate speech signals from MFCC features is described in [4]. The software used in this study was obtained from the Katholieke Universiteit Leuven and will be used in future experiments. HSR experiments based on this paradigm are to be conducted in the future.

## 7. CONCLUSIONS

A fair comparison of speech recognition of humans and machines has been carried out. Our results show that even if humans cannot exploit contextual knowledge, error rates of ASR systems are at least 300 % higher than human error rates. The results motivate a further investigation of to what extent principles of human hearing can be used as blueprint for ASR feature extraction.

An interesting analysis regarding consonant confusions has been carried out by Sroka and Braida [6], who compared recognition scores of humans and machines. Since we used the OLLO database for the man-machine-comparison, a wider range of speech intrinsic variabilities (like articulation characteristics and different dialects), a larger number of speakers and the recognition scores of consonants and vowels were taken into account for ASR and HSR experiments.

The results from speech intelligibility tests with human listeners suggest that consonant identification is a more difficult task for humans than recognition of vowels is. This becomes most obvious in noisy test conditions, where recognition accuracies are constantly better for CVCs that for VCVs.

An analysis of phonetic features shows that voicing is a feature crucial for discrimination of certain consonants. While human listeners seem to optimally exploit the voicing information, this feature is often misclassified by ASR systems, which leads to the most prominent errors in the consonant matrix of confusion. This suggests to use features for ASR where the voicing information is taken into account.

## 8. ACKNOWLEDGMENTS

The research reported in this paper was funded by the European Union in the Divines project (no. IST-2002-002034, URL: http://www.divines-project.org) and the SFB/TR31 'Das aktive Gehör' (URL: http://www.uni-oldenburg.de/sfb-tr31).

## 9. REFERENCES

[1] Wesker, T.; Meyer, B.; Wagener, K.; Anemüller, J.; Mertins, A. and Kollmeier, B. - "Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines", Interspeech 2005, pp. 1273-1276, 4-8 Sept. 2005, Lisbon, Portugal.

[2] Lippmann, Richard P. - "Speech recognition by machines and humans", Speech Communication 22, 1-15, 1997.

[3] Christiansen, Thomas U. and Greenberg, S. - "Frequency Selective Filtering of the Modulation Spectrum and its Impact on Consonant Identification.", "Hearing aid fitting" 21.st Danavox Symposium 31/8 - 2/9, 2005, Denmark.

[4] Demuynck, Kris; Garcia, Oscar and Van Compernolle, Dirk - "Synthesizing Speech from Speech Recognition Parameters",Proc. International Conference on Spoken Language Processing, volume II, pages 945–948, October 2004, Jeju Island, Korea.

[5] G.A. Miller, G.A.; Nicely, P.E. -"An Analysis of Perceptual Confusions Among Some English Consonants", JASA(2), 338-352, 1955.

[6] Sroka, J.; Barida, L. - "Human and Machine Consonant Recognition", Speech Communication 45 (401-423), 2005.