Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes

Bernd T. Meyer,[a] Thomas Brand, and Birger Kollmeier
Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany

(Received 13 November 2009; revised 13 October 2010; accepted 14 October 2010)

The aim of this study is to quantify the gap between the recognition performance of human listeners and an automatic speech recognition (ASR) system with special focus on intrinsic variations of speech, such as speaking rate and effort, altered pitch, and the presence of dialect and accent. Second, it is investigated if the most common ASR features contain all information required to recognize speech in noisy environments by using resynthesized ASR features in listening experiments. For the phoneme recognition task, the ASR system achieved the human performance level only when the signal-to-noise ratio (SNR) was increased by 15 dB, which is an estimate for the human–machine gap in terms of the SNR. The major part of this gap is attributed to the feature extraction stage, since human listeners achieve comparable recognition scores when the SNR difference between unaltered and resynthesized utterances is 10 dB. Intrinsic variabilities result in strong increases of error rates, both in human speech recognition (HSR) and ASR (with a relative increase of up to 120%). An analysis of phoneme duration and recognition rates indicates that human listeners are better able to identify temporal cues than the machine at low SNRs, which suggests incorporating information about the temporal dynamics of speech into ASR systems.

© *2011 Acoustical Society of America*. [DOI: 10.1121/1.3514525]

PACS number(s): 43.72.Ne, 43.71.Es, 43.72.Ja [SSN] Pages: 388–403

I. INTRODUCTION

While human listeners have little difficulties in dealing with recognition of spoken language in acoustically challenging situations, automatic speech recognition (ASR) often lacks the same robustness that is achieved by the auditory system. This observation has motivated research that compares the recognition performance of human speech recognition (HSR) and ASR systems, with the ultimate goal of improving automatic recognizers by learning from the principles in the human auditory system: For example, Lippmann (1997) reported that the gap between HSR and ASR scores (or the "human–machine gap") widens with a higher complexity of the recognition task, e.g., due to the addition of noise or an increase of the vocabulary size. For very complex tasks (such as the transcription of spontaneous speech) the error rates of ASR were reported to be an order of magnitude higher than those of HSR. In more recent studies that compare HSR and ASR scores, various aspects of the robustness against such *extrinsic* sources of variation (i.e., the variation that arises from factors which are not associated with the speech signal itself, such as additive or convolutional noise) have been covered in more detail. Human listeners were shown to outperform ASR systems in speech-like or modulated additive noise, both for a digit recognition task that avoids the use of grammatical context, as well as for a consonant recognition experiment based on nonsense syllables (Carey and Quang, 2005; Cooke and Scharenborg, 2008). Cooke and Scharenborg (2008) reported ASR baseline error rates to be 85% higher for clean speech compared to HSR error rates on the consonant recognition task. In another set of experiments, Sroka and Braida (2005) compared HSR and ASR scores for nonsense syllables and investigated the effect of additive noise and high- and low-pass filtered speech. The addition of noise resulted in considerable differences between HSR and ASR recognition scores, i.e., an ASR system using cepstral features reached human recognition performance only when the signal-to-noise ratio (SNR) was increased by 10 dB. On the other hand, high- and low-pass filtering reduced (and for some conditions even eliminated) the gap. Sroka and Braida conclude—based on an analysis of articulatory features (AFs)—that human listeners and automatic classifiers use different cues, which is supported by relatively low correct identification scores for the voicing feature.

While the robustness of ASR systems against extrinsic variability (especially when arising from additive noise) has been studied extensively, the robustness against *intrinsic* variations of speech (i.e., the natural variability that is produced by the talker) is far less understood. Factors that contribute to this intrinsic variability are foreign and regional accents, speaker physiology, speaking style and spontaneous speech, rate of speech, and the speaker's age and emotional state. Even though human listeners are remarkably robust in their recognition performance against these intrinsic variations, this does not apply to common ASR systems. The aforementioned variations were found to degrade the classification performance of automatic recognizers even when the acoustic conditions are optimal (Benzeguiba *et al.*, 2007). For example, the ASR recognition scores decrease when the rate of speech is changed compared to a normal speaking rate (Stern *et al.*, 1996). For human listeners, the dependency between speaking rate and recognition scores is far less pronounced (Krause and Braida, 1995).

[a] Author to whom correspondence should be addressed. Electronic mail: bernd.meyer@uni-oldenburg.de

In the present study, a comparison of human and ASR performance with special focus on *intrinsic* variations is performed, with the aim of quantifying the impact of such intrinsic factors on human and ASR performance. This comparison may be seen as a first step toward improving ASR systems by incorporating knowledge from the field of HSR (as, e.g., proposed by Scharenborg, 2007). The variabilities under consideration were speaking rate, speaking effort (i.e., loudly and softly spoken speech), speaking style (utterances with rising pitch), and dialect and accent (i.e., phonological differences compared to a standard language that depend on the regional origin of the speaker). The effect of these parameters on the recognition performance is compared to other factors that are known to influence the classification [such as the choice of speaker (Barker and Cooke, 2007)].

A second aim of this study is to quantify the contribution of feature extraction to the differences observed between the recognition performance in HSR and ASR. The underlying question is: Is the informational content of the most common ASR features sufficient to recognize distorted speech when using a nearly optimal classifier (i.e., a human listener)? Furthermore, the use of stimuli spoken with different speaking styles enables an evaluation of the interaction between feature extraction and intrinsic variations.

From a physiological point of view, the feature extraction stage may be interpreted as the bottom–up processing performed by the human auditory system (which is assumed to be mainly driven by the acoustic stimulus, so that the "internal representation" of that stimulus is created without feedback from higher stages in the auditory pathway). Similarly, the ASR back-end may be seen as the equivalent to the top–down component of the auditory system, since it generates a hypothesis of phoneme combinations to be considered in the recognition process, which implies a retroaction of the concurring hypotheses on word selection. In this physiological interpretation, the gap in phoneme recognition between man and machine has a bottom–up component (due to imperfect representation of speech by the acoustic features) and a top–down component (caused by imperfect classification techniques). This analogy between signal processing in HSR and ASR and the analysis of the individual processing steps may be of value for building models of human speech perception based on ASR techniques.

Our approach to assess the informational content of features is to convert mel-frequency cepstral coefficients (MFCCs) to audible signals (i.e., to resynthesize them) and present these signals to human listeners. The original signals were also used in listening tests to evaluate the HSR–ASR performance gap independently of the feature extraction stage. Listening experiments based on resynthesized speech have been conducted earlier. Leonard (1984) performed tests with clean digits that were resynthesized from linear prediction coefficients. The recognition accuracy based on the majority of three listeners was 99.9%, indicating that signals resynthesized from the spectral envelope of short-time fragments of speech are sufficient in acoustically optimal conditions. Peters *et al.* (1999) carried out a comparison of HSR and ASR recognition performance with unaltered and resynthesized speech. Feature vectors calculated from noisy digits were converted to audible signals based on an analytical processing scheme. When comparing HSR and ASR scores based on informational equivalent features, the digit error rate of ASR was found to be 13.1%, while the scores for the original and resynthesized features was 2.9% and 10.7%, respectively. However, the effect of a specific SNR level as well as speech-intrinsic variability is not investigated in these studies. For similar experimental conditions, the same speech database with nonsense syllables was employed for ASR and HSR tests. Using the same database has the advantage of suppressing unwanted variability that is, e.g., caused by interindividual differences across speakers. This might be especially important when investigating changes in speaking rate, as different talkers often employ different strategies to produce speech within speaking rates (Krause and Braida, 2003). One of the limitations of the current study is the restriction to nonsense combinations of phonemes. Results for conversational speech are expected to be different, since in this case humans profit from their sophisticated language model that can compensate for phoneme errors better than the language models applied in ASR (Shen *et al.*, 2008). Further, when continuous, meaningful speech is investigated, human listeners are able to exploit additional cues (e.g., due to knowledge about the context) that are often not available to ASR systems. Since the aim of this study is to perform a fair comparison of recognition performances, these effects were not considered by investigating the recognition of sublexical units.

II. METHODS

An overview of the HSR and ASR experiments and the most important experimental parameters are presented in Table I. Detailed experimental design is presented in the following subsections.

A. Speech database

The corpus used for this study is the Oldenburg Logatome (OLLO) corpus (Wesker *et al.*, 2005).[1] It consists of nonsense utterances (logatomes), which are composed according to phonetic and phonotactic rules. The logatomes are combinations of consonant-vowel-consonant (CVC) and vowel-consonant-vowel (VCV) with identical outer phonemes. The database is used to analyze the performance of human listeners in phoneme recognition where the task is to identify the middle phoneme, which limits the number of response alternatives and allows for an easy realization of HSR tests. This response format was chosen in order not to overload the subject with too many response alternatives (i.e., a maximum number of 14 alternatives is still manageable by trained subjects). In addition, keeping the initial and final vowel fixed mostly eliminates the subject's tendency to respond to a meaningless logatome with the closest known meaningful word. The phonemes contained in the database produce above-average phoneme error rates in either human or ASR. Phonemes resulting in high error rates are of special interest in this study, since the differences between HSR and ASR should be analyzed. Note that only phonemes that occur in the German language are considered, and the selection of phonemes was based on studies that analyze German words or sentences [for details, please

TABLE I. Overview of the experimental setup for the HSR and ASR experiments.

Aim of experiment: Analysis of the effect of speaking rate, effort, and style	
HSR type of signal	Original signals (SNR: −6.2 dB) Resynthesized signals (SNR: +3.8 dB)
HSR test set	Set *RES*: CVC and VCV utterances with two speaking rates (fast/slow), effort (loud/soft) and style (question/normal), and four talkers (2M, 2F) 3600 utterances (150 logatomes × 4 speakers × 6 variation styles)
HSR listening subjects	Six normal-hearing subjects (3M, 3F)
Type of noise	Stationary, speech-shaped noise (all HSR and ASR experiments)
ASR features	MFCC features calculated from clean and noisy signals
ASR noise	SNRs: −6.2 to 18.8 dB (matched training and testing SNR) + clean utterances
ASR test set	Set RES_{EXT} (same as HSR test set, with additional repetitions, ∼10 800 utterances)
ASR training set	Set RES_{TRAIN} (ND speech, speaker-independent training, 6 talkers, 3M, 3F, ∼16 200 utterances)
Aim of experiment: Analysis of the effect of dialect and accent	
HSR type of signal	Original signals (SNR: −6.2 dB) Resynthesized signals (SNR: +3.8 dB)
HSR test set	Set *DA*: CVC and VCV utterances with and without dialect/accent, normal speaking style, 10 talkers (5M, 5F) 1500 utterances (150 logatomes × 2 speakers per region × 5 regions)
HSR listening subjects	Five normal-hearing subjects (2M, 3F)
Type of noise	Stationary, speech-shaped noise (all HSR and ASR experiments)
ASR features	MFCC features calculated from clean and noisy signals
ASR noise	SNRs: −6.2 to 13.8 dB (matched training and testing SNR) + clean utterances
ASR test set	Set DA_{EXT} (same as HSR test set, with additional repetitions, ∼4500 utterances)
ASR training set	Set DA_{TRAIN} (normal speaking style, speaker-independent training, 40 talkers, 3M, 3F, ∼18 000 utterances)

refer to Wesker *et al.* (2005)]. Therefore, the results obtained with an English set of phonemes [as, e.g., presented in Sroka and Braida (2005) and Cooke and Scharenborg (2008)] may differ from the data obtained with the OLLO corpus. The phonemes contained in the database as well as other important properties are listed in Table II.

1. Sources of intrinsic variation and choice of speakers

The OLLO corpus employed in this study contains utterances with systematically varied intrinsic variabilities, which have been chosen based on ASR experiments that compared the performance of automatic recognizers with these variabilities present or not. The variabilities cover changes in speaking rate, effort, and style, as well as dialect and accent (Table I).

Ten speakers originating from the northern part of Germany (Oldenburg near Bremen and Hannover) were recorded. The spoken language in this region is usually considered as standard German (Kohler, 1995), which is therefore referred to as "no dialect" (ND). Furthermore, data from ten speakers from each of the following dialect regions was collected: From the northern part of Germany [East Frisian (EF) dialect], from East Phalia [East Phalian (EP) dialect] near Magdeburg, and from locations near Munich in Bavaria [Bavarian (BV)]. Accented speech was recorded in Mons in Belgium. The first language of the Belgian speakers was French (FR); they did not speak German but were supplied with an adapted transcription of the VCV and CVC utterances to ensure the desired realization of phonemes, which resembles the pronunciation of FR talkers producing German phonemes. Five female and five male speakers from each region were recorded, resulting in a total of 50 speakers. The age of subjects varied between 18 and 65 years. Each logatome was recorded in "neutral/clear" speaking style as a reference. In addition, one of the five selected variabilities (fast and slow speaking rate, loud and soft speaking style, and condition

TABLE II. Properties of the OLLO corpus.

Number of speakers	50 (25 male, 25 female)
Number of different VCVs	70 (14 central consonants combined with 5 outer vowels) Consonants: Fricatives (/f/, /s/, /ʃ/, /ts/, /v/), Plosives (/p/, /t/, /k/, /b/, /d/, /g/), Nasals (/n/, /m/), Lateral approximant (/l/). Vowels: /a/, /e/, /i/, /ɔ/
Number of different CVCs	80 (10 central vowels combined with 8 outer consonants) Consonants: Fricatives (/f/, /s/), Plosives (/p/, /t/, /k/, /b/, /d/, /g/). Vowels: /a/, /a:/, /ɛ/, /e/, /ɪ/, /i/, /ɔ/, /o/, /ʊ/, /u/
Number of different logatomes	150
Number of speaking styles	5 + reference condition (fast, slow, loud, soft, question, and normal)
Number of dialects/accents	4 + reference condition (EF, BV, EP, FR, and ND)
Utterances per speaker	2700 (150 logatomes × 3 repetitions × 6 speaking styles)

"question" which refers to rising pitch) was altered for each of the subsequent recordings. Note that the database does not include combinations of the variations of rate, effort, and style. To provide a broad test and training basis for ASR experiments each logatome was recorded three times which resulted in 150 logatomes × 6 speaking styles × 3 repetitions = 2700 utterances per speaker.

2. *Recording conditions and phonetic labeling*

Speakers for the OLLO database read a transcription of the CVC or VCV and were asked to produce the logatome in one of the six speaking styles. The transcription was prepared by a phonetician and represented on a computer screen by its most frequent grapheme combination. Different transcriptions for the German and FR speakers were used, since the interpretation of most grapheme combinations also differs for these groups. The speakers were supervised during the recordings and periodically reminded to speak in the desired manner. All VCV stimuli were produced with front stress, which corresponds to the common pronunciation of two-syllable German words. They were low-pass filtered with 8 kHz cutoff frequency and sampled down to 16 kHz, which is a typical sampling rate for many ASR tasks.

The OLLO corpus was phonetically time-labeled, i.e., temporal positions of phoneme boundaries have been determined for each utterance, making it suitable for tasks such as training of phoneme recognizers. Labeling was performed with the MUNICH AUTOMATIC SEGMENTATION SYSTEM (MAUS) software package provided by the Bavarian Archive for Speech Signals (BAS). The MAUS labeling is similar to forced alignment approaches based on hidden Markov models (HMMs). However, in contrast to standard forced alignment, it takes into account pronunciation variations typical to a given language by computing a statistically weighted graph of all likely pronunciation variants. The consideration of such variants is an important feature for logatomes spoken by speakers with dialect or accent. For example, the realization of the VCV /o ʃ o/ is [o ʃ o] for standard German, but [o ʃ æ] for the EP dialect. The MAUS software provides connected phonetic labels, i.e., closures of plosives are labeled as the corresponding phoneme. For details, the reader is referred to Kipp *et al.*, 1996. All 150 logatomes were transcribed in the speech assessment methods phonetic alphabet (SAMPA), and the transcription was used as input for the time-labeling procedure.

B. Preparation of speech stimuli

Speech intelligibility tests with human listeners included two conditions, i.e., the presentation of noisy (but otherwise unaltered) signals and listening tests with speech tokens that were resynthesized from ASR features. The latter were obtained by decoding the feature vectors used internally by the speech recognizer to acoustic speech signals.

Unaltered speech signals from the OLLO database are used to measure the phoneme recognition performance gap between HSR and ASR in the presence of intrinsic variabilities and additive noise. Stationary speech-shaped noise is added to the signals to prevent ceiling effects (cf. Sec. II C). A second experimental condition covers the aspect of resynthesized speech. The resynthesis of speech is based on the most common features in ASR, i.e., MFCCs. 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 the gender is usually not recognizable. Nevertheless, resynthesized speech items remain intelligible in the absence of noise (Demuynck *et al.*, 2004). To analyze if this holds true for noisy speech, the same type of masking noise employed for the original signals was added for resynthesized signals. By adding noise, redundant information in the speech signal is masked, so that intelligibility is potentially decreased. The reduction of redundancy might be particularly critical in the presence of speech-intrinsic variabilities.

1. *Calculation of cepstral coefficients*

MFCCs are a compact representation of speech signals and have been successfully applied to the problem of ASR (Davis and Mermelstein, 1980). This compact representation has been optimized to retain the information necessary for ASR, while information about speech quality and the individual speaker is mostly discarded. Specifically, the phase information and fine structure of the spectrum are disregarded. However, this may be detrimental in noisy conditions, because information such as fine phonetic detail which is exploited by humans for speech understanding (Hawkins, 2003) is masked in the noisy condition.

In order to calculate MFCC features from speech, signals with 16 kHz sampling frequency are windowed with 30 ms Hanning windows and a frame shift of 10 ms. Each frame undergoes the same processing steps: Calculation of the amplitude spectrum, reduction of the frequency resolution using a mel-scaled filter bank and calculating the logarithm, and the inverse discrete cosine transformation (IDCT) of its output. The 12 lowest coefficients plus an additional energy feature are selected for the ASR experiments and HSR tests with resynthesized speech. This results in (mostly decorrelated) cepstral coefficients, where lower coefficients characterize the coarse structure of the spectrum, while higher coefficients code the fine structure caused by the excitation of the vocal tract. These feature vectors were used for the ASR tests as well as the basis for resynthesized speech presented to human listeners. In ASR experiments, additional delta and acceleration coefficients were used. These are calculated directly from the cepstral features, i.e., no additional information is provided to the ASR system.

2. *Resynthesis of cepstral coefficients*

In order to reconstruct an acoustic speech signal from MFCC features, the spectral envelope has to be reconstructed from the feature data. This is done using a linear neural network that inverts the discrete cosine transformation. In this study, noisy training material from the OLLO training subset has been used to determine the optimal weights of the neural net. In a second step, the spectral fine structure and phase information has to be estimated. Since the listeners' knowledge should be limited to the information contained in the features, additional information such as voicing or fundamental frequency should not be added during the decoding process, i.e., an artificial excitation signal has to be used. This signal may be a pulse train (which corresponds to voiced excitation of the vocal tract), a noise signal (as for voiceless excitation), or a superposition of these. The artificial excitation signal is defined by the fundamental frequency and the amount of voicing. For high-quality resynthesis, these parameters need to be extracted from the speech signal. In this study, however, it would give human listeners an unfair advantage against the ASR system. Therefore, this type of information is not employed for the resynthesis.

The excitation signal $p(t)$ is combined with the smoothed magnitude spectrogram $|E(kT, \omega)|$ by calculating the dot product of $|E(kT, \omega)|$ and the magnitude spectrogram of $p(t)$ (i.e., $|P(kT, \omega)|$, where k is the frame index, T is the frame shift, and ω is the frequency tab. This leads to the target magnitude spectrogram $|Y(kT, \omega)|$ of the resynthesized signal. In order to construct the phase information, an algorithm proposed in (Griffin and Lim, 1984) is used. This algorithm iteratively decreases the squared error between $|Y(kT, \omega)|$ and the magnitude spectrogram $|X_i(kt, \omega)|$ of the resynthesized signal. At each iteration i, the next estimate of the time signal $x_{i+1}(t)$ is constructed from the target magnitude spectrum $|Y(kT, \omega)|$ combined with the phase spectrum of the previous estimate of the time signal $x_i(t)$. The algorithm usually converges in less than 100 iterations, even if white noise is used as the initial time signal $x_i(t)$ (Griffin and Lim, 1984; Demuynck *et al.*, 2004).

Since the properties of the excitation signal are crucial for the overall quality of resynthesis, preliminary tests were performed which showed that—in the presence of speech-shaped noise—intelligibility is higher when a pulse train is used as an excitation signal (instead of noise or a mixed noise-pulse signal), which is therefore used for all presented HSR tests with resynthesized speech. A fundamental frequency of 130 Hz was chosen for all presentations. Due to the fixed fundamental frequency, resynthesized speech sounds artificial and tinny, but remains understandable in the absence of noise. This algorithm was kindly supplied by the Katholieke Universiteit Leuven.

Pilot measurements with one listening subject and a reduced test set have shown that HSR scores are usually very close to 100% for the clean condition, both for the unaltered and the resynthesized signals. This confirms findings from Meyer and Wesker (2006), who found a recognition rate of 99.1% for non-dialect speech 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. Hence, a stationary noise signal with speech-like frequency characteristics was chosen as masker for the logatomes (Dreschler *et al.*, 2001). It was introduced by the International Collegium of Rehabilitative Audiology (ICRA) and implemented by adding artificial speech signals that represented a single speaker speaking with normal effort. The spectral and temporal properties were controlled and had a close resemblance to real-life communication without clear modulation. In case of resynthesized speech, noise is added *before* MFCCs are calculated from the original signals. This resembles the signal processing chain in ASR, in which features are extracted from noisy speech.

The pilot measurements also showed that *noisy* resynthesized speech resulted in considerably lower recognition scores compared to noisy original signals, when an identical SNR was used for both conditions. Depending on the value of the SNR, these differences either resulted in a ceiling of scores obtained with original signals or in flooring of scores measured with resynthesized utterances, which also aggravates the collection of statistically valid data. 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 a SNR of 3.8 and −6.2 dB, respectively.

C. HSR and ASR test and training sets

Two sets of utterances, which are subsets of the OLLO corpus, were defined to analyze the effects of speech-intrinsic variabilities on human and automatic recognition performance. The compilation of subsets was necessary since the presentation of all utterances in the database (with more than 120 000 individual recordings) is not tractable in listening experiments. The total number of speakers was chosen with regard to the resulting size of each set. The effects due to changes in speaking rate, effort, and style were analyzed using Set rate, effort, and style (*RES*), which contained data from four talkers (without regional dialect or accent) spoken in six variabilities. The influence of dialect and accent was analyzed with Set dialect and accent (*DA*), which contains normally spoken utterances from two speakers from each dialect/accent region. The speakers' gender was balanced in both sets and for each dialect region. The properties of the test sets are summarized in Table I.

1. Speaker selection

Since earlier studies have shown that the intelligibility of speech strongly depends on the choice of speaker (Barker and Cooke, 2007), a selection procedure was carried out to avoid the use of speaker data that produce very high or low scores compared to the complete data set. In order to find speaker sets that are representative for the complete database, a standard ASR system was trained with all utterances from 49 speakers and tested with the speech data of the remaining speaker. The ASR system used MFCC features and a HMM classifier, which were configured as described in Sec. II D 2. This procedure was performed for all speakers in the corpus; the corresponding scores for each speaker are presented in Table III. The four speakers selected for Set *RES* (framed elements in Table III) produced an average score in the same range as the average scores measures for all speakers without dialect (84.3% vs 84.1%). Similarly, the scores often speakers included in Set *DA* (bold elements in Table III) were comparable to the score averaged over all 50 speakers (75.4% vs 75.8%, respectively).

2. Calculation of the SNR

For the measurements with Set *RES*, the SNR was calculated by relating the root-mean-square (rms) value of each

TABLE III. Average ASR recognition rates in percent for each of the 50 speakers in the OLLO database, which were obtained by training the ASR system with 49 speakers and testing using data from the remaining speaker. The speaker index as defined in the OLLO documentation is given by 10 × D + *S*. Scores of speakers selected for Set *DA* are printed in bold. For Set *RES*, utterances of speakers 1, 2, 6, and 8 without dialect were selected. The average score over all 50 speakers is 75.8%.

	Number of speakers (S)										
Dialect/accent	1	2	3	4	5	6	7	8	9	10	Average
ND (D = 0)	**80.8**	**88.4**	86.5	87.1	79.9	81.0	79.0	86.8	84.3	86.8	84.1
EF (D = 1)	69.4	78.8	83.9	77.1	75.6	88.7	**82.6**	77.6	**76.3**	86.8	79.7
EP (D = 2)	79.9	66.9	**76.9**	84.5	69.4	69.2	86.0	67.1	79.9	**72.2**	75.2
BV (D = 3)	85.4	**70.3**	67.0	55.2	83.5	78.8	69.9	81.9	60.4	**71.5**	72.4
FR (D = 4)	**65.6**	**69.3**	74.7	66.3	79.9	79.6	59.2	60.6	57.6	63.6	67.6

individual signal (*including silence*) and the rms value of a stationary masking noise of equal length. The speech files in Set *RES* contain a silence of 500 ms duration at the beginning and end of each signal. However, for the utterances in Set *DA*, we observed variations in the duration of the silence segment. Therefore, the application of the aforementioned scheme of SNR calculation resulted in higher intelligibility for utterances with short silence segments and lower intelligibility for logatomes with longer silence segments. Hence, a different SNR calculation scheme was applied for utterances in Set *DA*, based on the rms values of the speech segments (*excluding* silence) of each single audio signal and the rms value of a stationary masking noise of equal length. A simple voice detection algorithm based on an energy criterion was used to identify the beginning and the end of the logatome in the audio file. Random control samples were chosen to control the proper functioning of the algorithm. Since the length of silence before and after each logatome is 500 ms for Set *RES* and because the variation of temporal spread of identical logatomes is relatively small, the two calculation schemes result in a fixed offset which was found to be 3.8 dB compared to the SNR calculation scheme mentioned above. For clarity, the SNR values for Set *RES* are converted to the first mentioned method.

3. ASR test and training sets

ASR testing was performed with the utterances from Set *RES* and Set *DA*, which are also used for the HSR experiments (Table I). The two additional repetitions recorded for OLLO which are not contained in the HSR test set were also included; this extended set is therefore referred to as Set RES_{EXT}. The extension violates the rule of having exactly equal conditions for HSR and ASR but increases the amount of test data by a factor of 3 at the same time. Since speakers were recorded in one session, the differences between utterances are expected to be negligible. Moreover, from the three recordings of each logatome in each variability an arbitrary file has been chosen for the HSR test, which prevents a systematic error when using three recordings instead of one. Finally, in HSR, the available number of individual responses is increased by collecting data from several listening subjects. On the other hand, the ASR system only produces a single decision for each utterance. When phoneme recognition scores are compared, the increased number of utterances in the extended test set outweighs the differences between the original and the extended sets because of reasons of statistical relevance.

ASR training was carried out with utterances from speakers not included in the test sets, resulting in speaker-independent recognizers. The properties of the training and test set are presented in Table I. The aim of the experiments carried out with Set *RES* was to analyze the effect of rate, effort, and speaking style on the recognition performance; hence, dialectal speech was not included in the ASR training and test sets. Analogously, the training set for experiments with Set *DA* included data from the remaining dialect speakers (40 speakers) whose data were not included in Set *DA*. The phonemes, gender, and the systematically varied parameters were equally distributed in the training and test sets.

ASR recognition scores were obtained for different SNRs, using the same masking noise employed for the HSR measurements. Since the scope of this study lays on *intrinsic* variations, we chose a matched training/testing paradigm for ASR to reduce the effect of SNR mismatch (which would introduce additional extrinsic variation between training and test). Signals with SNRs ranging from −6.2 to 18.8 dB, as well as clean signals, were used for training and testing in order to cover a wide range from very low to high ASR recognition scores.

D. Experimental setup

1. Tests with human listeners

Six normal-hearing listeners (three male and three female) without noticeable regional dialect participated in the listening tests for Set *RES*; five of these subjects (two male and three female) also participated in tests with Set *DA*. The listeners' hearing loss did not exceed +10 dB at more than two frequencies in the pure tone audiogram (with the exception of one subject, whose hearing loss was 15 dB at 8 kHz on one ear). Signals were presented in a soundproof booth via audiological headphones (HDA200, Sennheiser corporation, Wedemark-Wennebostel, Germany). An online freefield equalization and pseudo-randomization of logatomes was performed by a modified version of the "Oldenburg Measurement Applications" which was run on a standard personal computer (PC). 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. In order to limit the influence of training effects, the listeners were presented a training sequence consisting of 150 logatomes on each measurement day; the completion of this task took about 5–8 minutes. The training sessions were identical to the actual measurements, i.e., no feedback was given. After the training session, subjects were presented a sequence of logatomes at a comfortable listening level of 70 dB sound pressure level (SPL), and the effect of speech level that is expected to influence the recognition of, e.g., softly and loudly spoken utterances, was compensated for.

For each presentation, the listening subject selected a logatome 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 cumulative measurement time for each subject was approximately 22 hours, including pauses and instructions for listeners. It was distributed across different days (including a daily training session prior to data recording) in order not to exceed three hours of measurement for each day and subject.

2. ASR test setup

ASR experiments were carried out with a HMM with three states (corresponding to one phoneme) and eight Gaussian mixtures per HMM state. The system was set up to 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 grouping utterances with the same outer phonemes and subsequently using each group to train and test the back-end.

The same MFCC features have been used for the ASR test as for the resynthesized signals in HSR experiments. Additional delta and acceleration features were added to the 13 cepstral coefficients, yielding a 39-dimensional feature vector per time step. Without these features, ASR performance would drop dramatically, because the HMM is not capable of modeling all dynamic aspects of speech as well as humans can.

E. Evaluation methods

In addition to the recognition scores, we report the relative transmitted information, T_r, of AFs with the aim of gaining insight into the nature of phoneme confusions in HSR and ASR. The AFs under consideration are place and manner of articulation and voicing (cf. Table IV). Vowel and consonant recognition scores are also reported in terms of the transmitted information, which is a measure that corrects for chance performance and hence allows for a direct comparison of consonant and vowel recognition scores.

TABLE IV. AFs, their feature values, and the phonemes that correspond to a specific feature value (based on the International Phonetic Alphabet).

Place	Bilabial (/p/, /b/, /m/), Labiodental (/f/, /v/), Alveolar (/t/, /d/, /n/, /s/, /ts/, /l/), Palato-Alveolar (/ʃ/), Velar (/k/, /g/)
Manner	Plosive (/p/, /t/, /k/, /b/, /d/, /g/), Nasal (/n/, /m/), Fricative (/s/, /f/, /v/, /ʃ/, /ts/), Lateral Approximant (/l/)
Voicing	Voiced (/b/, /d/, /g/, /v/, /n/, /m/ /l/), Unvoiced (/p/, /t/, /k/, /s/, /f/, /ʃ/, /ts/)
Backness	Back (/ɔ/, ʊ, /o/, /u/), Front (/a/, /ɛ/, /ɪ/, /a:/, /e/, /i/)
Height	Closed (/ɪ/, /ɔ/, /i/, /u/), Close-mid (/e/, /o/), Open-mid (/ɛ/, /ɔ/), Open (/a/, /a:/)

A question that arises when measuring recognition scores in the presence of intrinsic variations is: Which acoustic properties of the stimuli result in the observed scores? Hillenbrand *et al.* (1995) have shown that the duration of phonemes is an important cue in HSR phoneme recognition. Hence, the relation of phoneme duration and recognition scores in HSR and ASR was analyzed with the aim of investigating to what extent temporal cues are employed by humans and machines.

1. AFs and transmitted information

The relative transmitted information, T_r, of AFs (Miller and Nicely, 1955) is a measure of how well the input variable is transmitted or—in terms of speech recognition—how well a specific AF is recognized. The *absolute* transmitted information is calculated by $T(x,y) = -\sum_{i,j} p_{ij} \log(p_i p_j / p_{ij})$, with p_i and p_j denoting the *a priori* and *a posteriori* probabilities for the stimuli, respectively, and p_{ij} denoting a matrix element of the confusion matrix. The relative transmitted information is given by $T_r(x, y) = T(x, y)/H(x)$ with the source entropy $H(x) = \sum_i p_i \log(p_i)$. The AFs under consideration and their respective feature values are presented in Table IV. For each AF, a confusion matrix is derived from the phoneme confusion matrix by grouping the matrix elements that correspond to the recognition or misclassification of a certain feature value (e.g., all elements that correspond to the presentation of an unvoiced sound, when a voiced sound was classified). Furthermore, the recognition of consonants and vowels is analyzed by calculating the T_r scores from the respective confusion matrix.

2. Phoneme duration

In order to analyze the effect of phoneme duration on the recognition rates, the distribution of the duration of each middle phoneme was determined in a first step. The duration was derived from the phoneme boundaries that were automatically estimated using a modified forced alignment algorithm (cf. Sec. II A 2). Second, the durations of phonemes were grouped based on histograms, and the corresponding recognition rates for each bin were calculated. When too few bins for the histograms are used, the temporal resolution is suboptimal, while too many bins result in a low number of elements per bin and unreliable observations. We found 10 bins for the histograms to be a good compromise. Additionally, only bins with at least 50 individual decisions by human listeners were considered for the analysis. Since the ASR test set contained only half the number of items (cf. Table I), this threshold was set to 25 for the ASR data.

III. RESULTS

A. Overall performance

Overall HSR and ASR phoneme recognition scores obtained with Set *RES* and Set *DA* are presented in Table V.

TABLE V. HSR and ASR phoneme recognition scores in percent, depending on speech-intrinsic variabilities and the SNR. Scores for varied speaking effort, rate, and style were obtained with Set *RES*; the results for dialects and accents are based on measurements with Set *DA* (cf. Table I). The average over intrinsic conditions (first column) is broken down into vowel and consonant recognition score. For intrinsic variations, vowel and consonant scoring is presented in terms of the transmitted information in Sec. III B.

		Average	Consonants	Vowels	Normal	Fast	Slow	Loud	Soft	Question
HSR	Resynthesized (3.8 dB)	72.4	74.5	70.7	76.3	68.0	77.7	68.8	68.5	75.3
	Original (−6.2 dB)	74.5	67.7	80.5	78.6	72.3	77.2	79.3	63.3	76.3
ASR	Clean	80.4	85.2	76.3	85.3	78.8	82.3	76.6	79.1	80.5
	18.8 dB	77.5	78.9	76.3	83.5	76.1	81.3	73.1	75.9	75.3
	13.8 dB	76.0	76.4	75.6	82.3	72.9	80.7	71.3	73.1	75.6
	8.8 dB	72.8	69.8	75.4	80.4	67.9	78.4	67.0	69.5	73.8
	6.8 dB	69.7	65.2	73.6	76.3	64.5	75.8	65.2	65.8	70.6
	3.8 dB	64.5	56.4	71.5	71.5	60.0	69.5	60.3	59.2	66.2
	−1.2 dB	54.0	43.0	63.7	60.3	50.5	58.4	54.4	47.0	53.6
	−6.2 dB	31.8	22.1	40.2	35.2	35.5	32.2	36.2	21.0	30.5
		Average	Consonants	Vowels	ND	EF	BV	EP	FR	
HSR	Resynthesized (3.8 dB)	73.8	74.3	73.3	77.5	79.2	75.1	71.3	65.7	
	Original (−6.2 dB)	74.0	65.7	81.3	81.5	80.9	77.6	70.2	59.7	
ASR	Clean	82.1	85.1	79.4	88.4	84.5	79.1	84.1	74.2	
	13.8 dB	79.3	80.2	78.5	87.0	82.5	75.4	78.5	73.2	
	6.8 dB	73.6	68.3	78.3	81.5	76.9	71.5	72.4	65.9	
	3.8 dB	68.5	59.5	76.4	75.4	72.4	66.8	68.0	59.8	
	−6.2 dB	34.0	21.6	44.9	42.6	34.0	36.2	30.8	26.3	

ASR experiments were carried out at various SNRs, while the tests with human listeners were limited to a specific SNR level. When conditions with the same SNR are compared, strong differences between HSR with the original signals and ASR are observed. For utterances in Set *RES* (−6.2 dB SNR), the difference between the HSR score (74.5%) and the ASR result obtained at the same SNR (31.8%) amounts to 43% absolute. This corresponds to a relative increase of 168%. For Set *DA*, the HSR and ASR scores are 74% and 34%, respectively, resulting in an absolute difference of 40% or a relative increase of the phoneme error rate of 153%. The largest differences are observed for logatomes spoken with rising pitch ("question") and EF dialect, with differences of over 45% points. The smallest differences were found for speech with a FR accent and high speaking rate (with absolute differences of 33.4% and 36.8%, respectively).

A comparable average HSR performance with Set *RES* is obtained for the original and resynthesized signals (74.5% and 72.4% recognition rate, respectively). The information loss induced by the feature calculation and resynthesis can therefore be approximated and amounts to 10 dB [i.e., the SNR difference for original signals (−6.2 dB) and resynthesized signals (+3.8 dB)]. A comparable overall ASR performance (72.8%) is obtained at 8.8 dB SNR. Based on this observation, the overall gap for this phoneme recognition task to HSR performance can be estimated to be 15 dB. Results obtained with Set *DA* are consistent with these observations since the average HSR recognition scores obtained with original and resynthesized signals differ by only 0.2% absolute. At a SNR of 6.8 dB, the ASR recognition performance (73.6%) lies between these scores, indicating that the overall gap in terms of SNR amounts to 13 dB. For some conditions (soft and slow speaking style, EP dialect, and FR accent), the recognition rates for resynthesized signals are higher than those obtained with original signals, which can be attributed to the different SNRs that were used for these conditions.

Most intrinsic variations degrade HSR performance compared to the reference condition; the only improvements are observed for slow speaking style and EF dialect for resynthesized signals, loud speaking style for original signals, and categories "fast" and "loud" for ASR scores at the highest masker level.

1. Relative increase of phoneme error rates

The relative increase of errors in the presence of intrinsic variability for both HSR conditions and selected ASR experiments is displayed in Fig. 1. Results are presented for ASR experiments for which the same SNR as for the HSR measurements was used (−6.2 and 3.8 dB) and for normally spoken logatomes that resulted in comparable ASR performance (SNR +8.8 dB). For HSR, the respective error rate for normal utterances has been used as reference for the relative increase. The ASR reference is the error rate for normally spoken utterances at a SNR of 8.8 dB.

The largest differences between HSR with resynthesized and original signals are observed for loud and soft speaking style while all other conditions appear to be similarly influenced. The results for the ASR conditions with medium scores (SNR of 3.8 and 8.8 dB) are consistent, i.e., the highest degradations are observed for the conditions fast, loud, and soft.

In the presence of dialect and accent, the error rates of both HSR conditions increase (in the order, EF, BV, EP, and FR). When listening to the original, dialectal speech, the human error rates are up to 120% higher compared to the reference condition. For ASR, similar results were obtained, with the exception that BV results in slightly increased errors compared to EP.

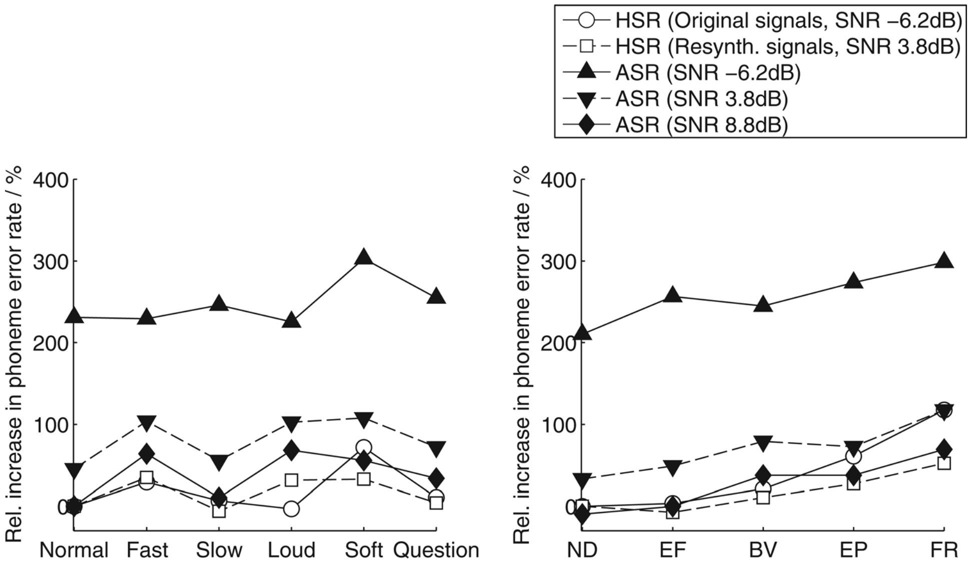

FIG. 1. Relative increase of phoneme error rates for HSR and ASR. The increase is related to the error rate obtained with normally spoken utterances for HSR. All ASR scores are related to normally spoken utterances with a training/test SNR of +8.8 dB (which produced similar performance compared to HSR). The scores are related to speaking rate, effort, and style (left panel), as well as dialect and accent. The categorical variabilities are depicted as connected line graphs for reasons of readability.

2. *Importance of speaking style, choice of speaker, and listener in phoneme recognition*

The importance of several parameters that may affect the HSR and ASR recognition rates was tested based on an analysis of variance. The explanatory parameters were speaking style (rate, effort, and rising pitch), choice of speaker, and choice of listener. The latter was included in the analysis of HSR results only. This choice of independent variables (IVs) resulted in 144 observations for HSR (6 speaking styles × 6 listeners × 4 speakers) and 24 observations for ASR (6 speaking styles × 4 speakers). The analysis demonstrated that all of these explanatory variables have a significant impact on the overall recognition performance in HSR and ASR (Table VI). Variations of speaking style appear to have a strong effect on the recognition results, both for HSR with original signals and for the ASR system. There are however shifts in the importance between the parameters "speaking style" and "choice of speaker," which are discussed in Sec. IV C.

The analysis was repeated for results from Set *DA*. Since the speaker and dialect/accent are mutually dependent variables, the analysis was limited to the IVs "dialect/accent" and "choice of listener" for HSR and to dialect/accent for ASR. This resulted in 50 individual observations for HSR (10 speakers × 5 listeners) and only 10 observations for ASR. In this case, significant results were only obtained for the parameter dialect in HSR and for ASR at the lowest masking level (Table VII). In order to rule out the possibility that insignificant results were obtained due to the low number of individual observations in ASR, McNemar's test was applied which is based on the individual decisions of the classifier and which has been proposed to test the significance of ASR results (Gillick and Cox, 1989). The test showed that in the low noise conditions, the results obtained with dialectal speech significantly differed from results for utterances spoken without dialect ($p < 0.001$). In most cases, the pair-wise comparison of dialect categories also showed significant differences between ASR recognition scores. We analyzed all combinations of dialect and accent in the database (i.e., combinations of ND, EF, BV, EP, and FR). From the 50 paired tests (5 SNRs × 10 pairs of dialects), only the following combinations were not significantly different: (ND and EF) (at −1.2, 3.8, and 8.8 dB SNR), (EF and BV) and (EF and EP) (both at −6.2 and 1.2 dB SNR), (BV and EP) (at 3.8 and 13.8 dB), and (BV and FR) (at 13.8 dB).

3. *Training effect*

A linear fitting of measurement results for human listeners shows the presence of a small training effect despite the training runs that each subject performed before actually

TABLE VI. Results of an ANOVA of HSR and ASR recognition scores obtained with Set *RES*. Speaking style relates to the categories of Set *RES*, i.e., fast, slow, loud, soft, question, and normal.

		Speaking style ($df = 5$)			Choice of speaker ($df = 3$)			Choice of listener ($df = 5$)		
		F	η^2	*P*	*F*	η^2	*P*	*F*	η^2	*P*
HSR	Original (SNR −6.2 dB)	23.4	28.4	<0.001	45.4	33.1	<0.001	5.8	7.0	<0.001
	Resynthesized (SNR 3.8 dB)	15.4	9.2	<0.001	189.1	67.7	<0.001	12.8	7.7	<0.001
ASR	−6.2 dB	6.2	50.8	<0.001	5.0	24.7	<0.05	—	—	—
	−1.2 dB	4.5	35.4	<0.05	8.6	40.9	<0.001	—	—	—
	3.8 dB	3.5	28.8	<0.05	9.6	46.9	<0.001	—	—	—
	8.8 dB	5.4	32.9	<0.001	13.3	48.8	<0.001	—	—	—
	13.8 dB	3.7	27.3	<0.05	11.2	50.3	<0.001	—	—	—

TABLE VII. Results of an ANOVA of HSR and ASR recognition scores obtained with Set *DA*.

		Dialect ($df = 4$)			Choice of listener ($df = 5$)		
		F	η^2	P	F	η^2	P
HSR	Original (SNR −6.2 dB)	22.4	66.5	<0.001	1.1	3.2	0.4
	Resynthesized (SNR 3.8 dB)	3.8	25.5	0.01	0.8	5.7	0.5
ASR	−6.2 dB	7.8	86.2	<0.05	—	—	—
	−1.2 dB	3.0	70.3	0.13	—	—	—
	3.8 dB	1.5	54.8	0.32	—	—	—
	8.8 dB	1.1	46.3	0.46	—	—	—
	13.8 dB	1.4	53.5	0.34	—	—	—

being tested. For experiments with the original signals, the average improvement is 0.9% between the first and the last of the 24 presented lists. In case of resynthesized signals, the training effect is larger (5.2% absolute improvement), possibly because the acoustic stimuli for this condition were unfamiliar to the listeners. Since the sequence of logatomes and test lists were randomized, no systematic error is expected from the training effect.

B. Transmitted information

The confusion matrices for consonants, vowels, and several AFs were used to calculate the relative transmitted information T_r associated with these features, as described in Sec. II E 1. Figure 2 presents the scores in relation to speaking style, rate and effort, and dialect and accent for both HSR and selected ASR conditions. The analysis based on AFs shows that conditions with comparable average performance (both HSR conditions and ASR at +8.8 dB SNR) exhibit considerable variations, both among different AFs and intrinsic variations. Scores obtained with resynthesized signals are in most cases higher than scores for original logatomes for consonant and consonant-associated features (left and center panels in Fig. 2). This may be attributed to the SNR difference between the two conditions (−6.2 and +3.8 dB). For vowel-associated features (right panels), the opposite trend is observed.

High speaking effort (loud) yields above-average performance for original signals, which can mainly be attributed to the voicing and place feature. This is however not observed for resynthesized features, for which only medium transmission scores are obtained. The overall low performance for original signals and low speaking effort (soft) is reflected in the T_r scores for all AFs, with the exception of voicing. For ASR, a strong dependence of the variability is observed for voicing. While normally and slowly spoken logatomes result in relatively high values for this feature, it is strongly degraded for the categories fast and loud (with degradations of 36% and 53% compared to the reference condition “normal”).

C. Phoneme duration

An analysis of phoneme durations was performed with the aim of identifying the phonetic–acoustic cues that result in the observed recognition rates of human listeners and the ASR system. The analysis showed that high and low speaking rates result in significant changes of phoneme durations, whereas the other speaking styles from Set *RES* (categories loud/soft/question) did not have a significant effect. Changes in phoneme duration are reflected in the average values and in the 5%/95% quantiles for variabilities fast (with 103 ms duration in average and 40/200 ms for the quantiles), slow (average: 255 ms, quantiles: 70/550 ms), and normal (average: 146 ms, quantiles: 45/316 ms).

The relationship between duration and recognition rate was analyzed on phoneme level for HSR and ASR scores (cf. Sec. II E 2). For consonants (Fig. 3) a clear relationship between duration and recognition score was not observed. In HSR, the fricatives /ts/ and /f/ yielded improvements with increasing duration, whereas for nasals the opposite relationship was found. Similarly, for the ASR condition (which resulted in comparable scores at 8.8 dB SNR) specific trends were not found.

On the other hand, for vowel identification in HSR (top left panel in Fig. 4) two clear trends can be identified: For the first group of vowels (D1 = /a/, /e/, /i/, /o/, /u/, open symbols in Fig. 4), the recognition rate of that phoneme decreases with increasing duration. The opposite tendency is observed for the second group of vowels (D2 = /a:/, /ɛ/, /ɪ/, /ɔ/, /ʊ/, closed symbols in Fig. 4). When produced at normal speaking rate, the vowels from group D2 exhibit a longer duration than those of Dl (Hillenbrand *et al.*, 1995); the decrease of scores of group Dl therefore appears plausible.

This trend for groups D1 and D2 can also be found for the resynthesized signals (top right panel in Fig. 4); however, for some examples (/ɛ/, /ɪ/) this result was not observed. The HSR recognition curves for D1 and D2 intersect at a duration of approximately 170 ms which therefore can be considered as an estimate of the category boundary between short and long vowels in HSR. The confusions of phonemes with durations comparable to this boundary (140–200 ms) were subject of further analysis. Utterances from the categories fast, slow, and normal were considered for this analysis. The following confusions produced the highest error rates (where the first and second phonemes correspond to the presented and chosen item, respectively): (/a/, /a:/), (/ɛ/, /e/), (/e/, /ɪ/), (/ɪ/, /e/), (/i/, /ɪ/), (/ɔ/, /o/), (/o/, /ʊ/), (/ʊ/, /o/), (/u/, /ʊ/). Hence, the phonemes can be pooled in three confusion groups (CG) that contain often confused vowels: CG_1 = (/a/, /a:/), CG_2 = (/ɛ/, /e/, /ɪ/, /i/), and CG_3 (/ɔ/, /o/, /ʊ/, /u/). Comparable error patterns were observed for HSR with resynthesized signals, with the

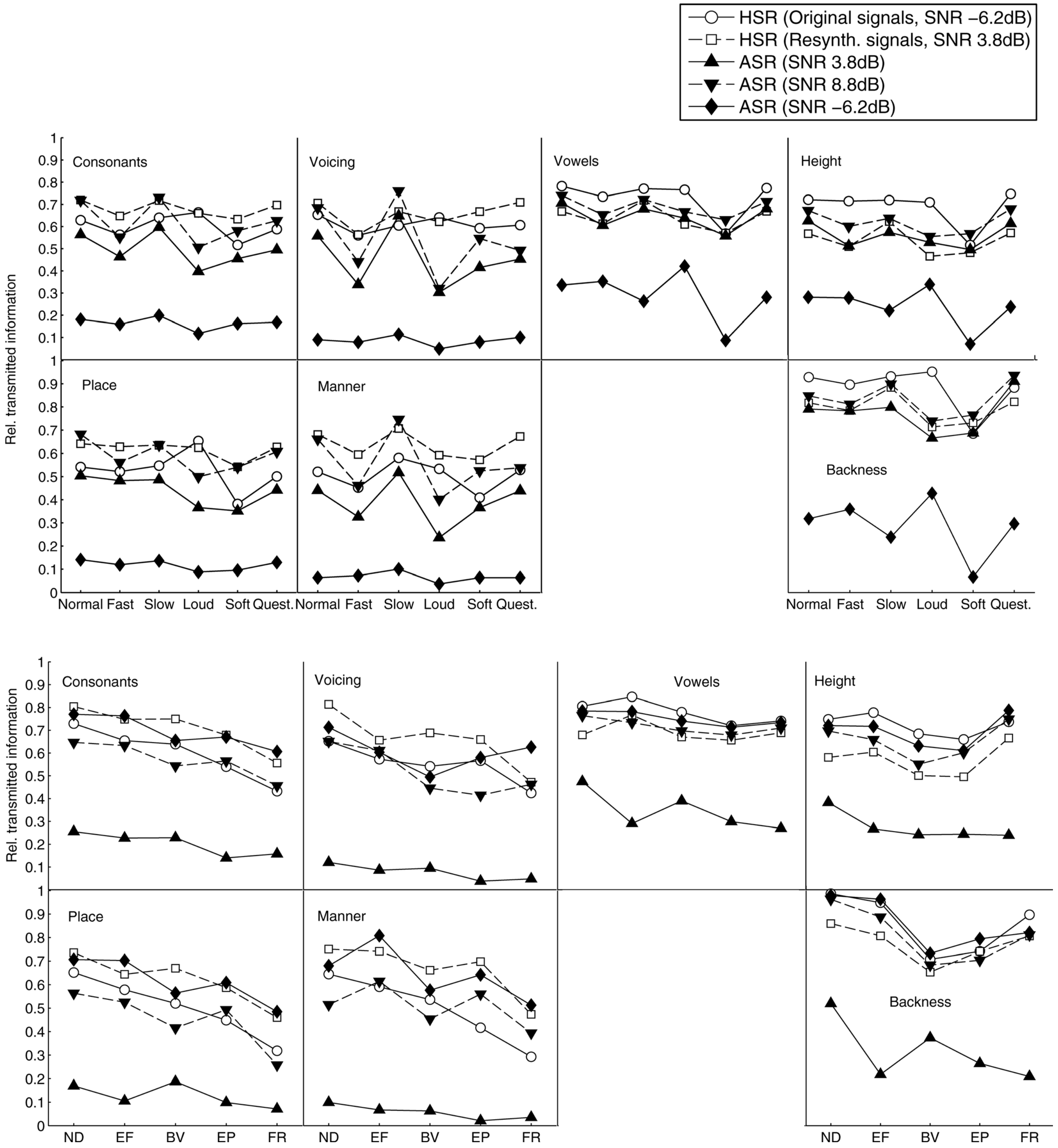

FIG. 2. Relative transmitted information scores for consonant, vowels, and AFs (cf. Table IV). The figure shows the influence of different speaking rates, efforts, and styles (upper panels) and the impact of various dialects.

exception of (/ɔ/, /a:/) that corresponds to an inter-group confusion with a high error rate. The corresponding vowel confusion matrices are shown as inlays in Fig. 4. For ASR at high SNRs, comparable overall trends were found, i.e., phonemes in D1 and D2 were similarly affected by the duration, and errors were mainly restricted to the same CGs (with the exception of (/a:/, /ɔ/) which produced high errors for all SNRs in ASR). However, at low SNRs, the relation between recognition and duration is not as pronounced as in HSR, since a reduced duration does not consistently result in an increased recognition of vowels in D1. For example, the scores for /ɔ/ and /ʊ/ decrease with duration while in the case of other phonemes (/o/ and /e/) a consistent trend is not observed at all. The errors between CGs can be obtained by averaging over the corresponding matrix elements in the confusion matrix. In ASR at low SNRs, many confusions occur between (CG_1 and CG_3) with an error rate of 33% compared to 1% in HSR at the same SNR and (CG_2 and CG_3) with 31%

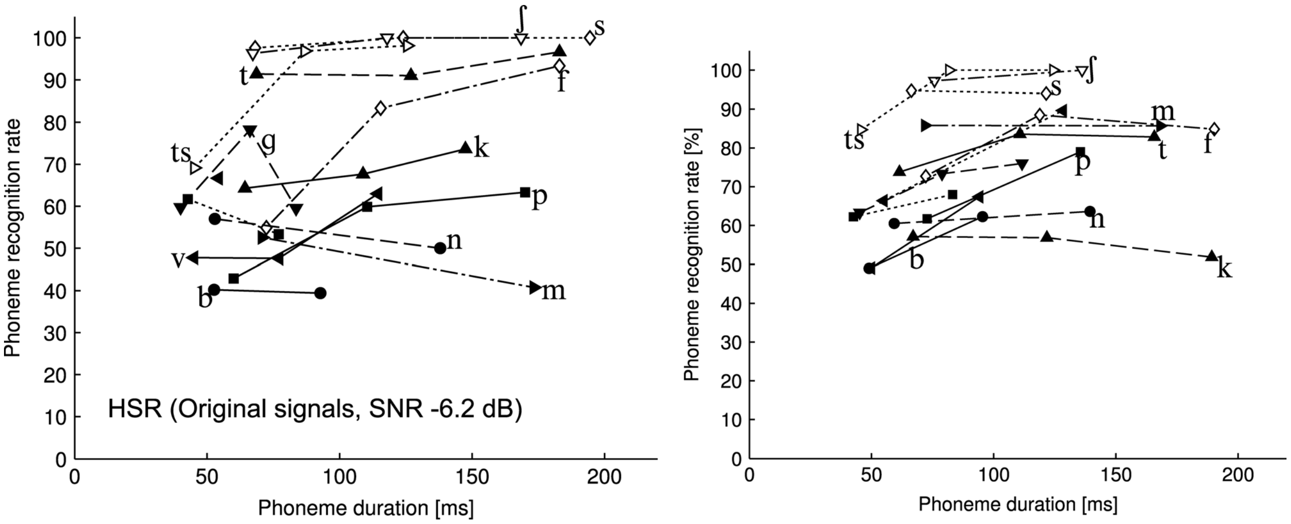

FIG. 3. Consonant recognition rates for Set *RES* in relation to phoneme duration. HSR scores obtained with original signals are compared to the ASR condition that produced comparable average scores (SNR 8.8 dB).

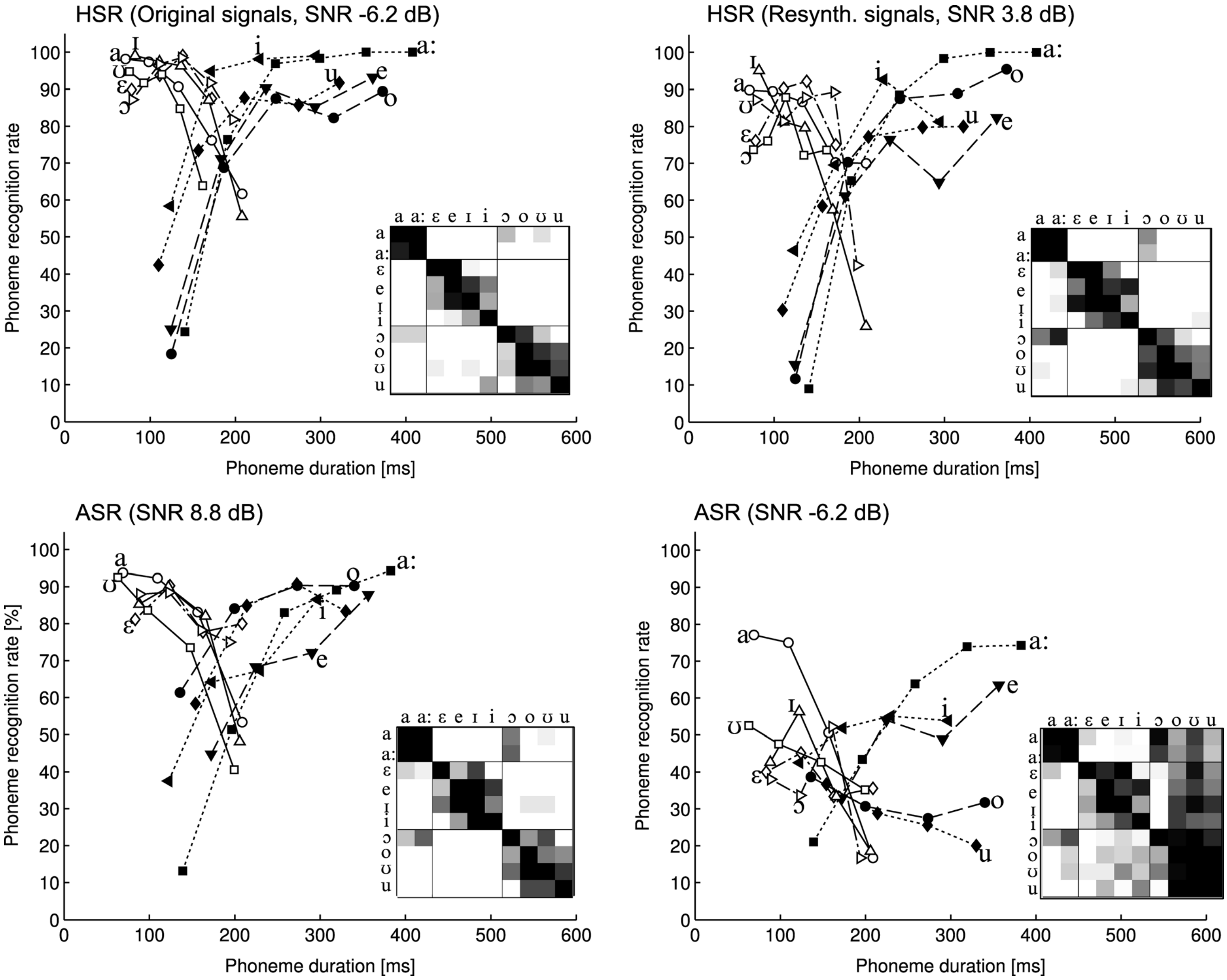

FIG. 4. Relation between phoneme duration and recognition rate for vowel phonemes (two HSR and two ASR conditions). The inlays show the confusion matrix of vowels with durations of 140–200 ms. The gray-scale intensity is proportional to the log of the value of each entry in the row-normalized CM, with black color representing unity and white color representing zero elements. Both HSR conditions are shown, as well as the ASR results obtained at 8.8 dB SNR (which resulted in a performance level comparable to HSR) and at −6.2 dB SNR (which has been used for HSR experiments with original signals).

compared to 0.1% in HSR. These errors are highly asymmetric, as the confusions (CG_3 and CG_1) and (CG_2, CG_3) exhibited error rates of only 5%.

IV. DISCUSSION

A. Human vs machine performance

A direct comparison of human and ASR performance showed that average phoneme scores in HSR are superior to the results obtained with a standard ASR system. One of the aims of this study was to quantify the gap in recognition performance of man and machine. By comparing the results obtained in HSR tests with original signals and the ASR recognition performance at identical SNRs, the human–machine gap can be expressed in terms of degraded recognition scores: For both test sets, the ASR scores were about 40% points lower than HSR scores. This corresponds to more than 150% relative increase of the phoneme error rate. Since the ASR experiments were carried out at several SNRs, the gap may also be quantified in terms of the SNR: To achieve the human performance level with Set *RES* (74.5% at −6.2 dB), the SNR used in ASR has to be increased to 8.8 dB, i.e., the SNR difference amounts to 15 dB. This is consistent with results obtained with Set *DA*, for which a gap of 13 dB was estimated. The machine results were obtained with a recognition system based on MFCCs combined with a HMM due to its widespread use in ASR. It can be expected that using other system architectures will result in different estimates of the human–machine gap.

These results can be compared to HSR and ASR scores from other studies. Lippmann (1997) reported an increase of word error rate (WER) by 400% for the automatic recognition of alphabet letters (based on the classification with a neural net). Cooke and Scharenborg (2008) used a VCV database to measure HSR scores as well as the ASR performance (based on MFCC features and an HMM classifier) and found a relative increase of 85% of error rates. In both studies, these results were obtained with clean speech (in contrast to noisy speech employed in this study), which may explain for the differences to the presented experiments. Additional factors that may contribute to these differences are the different phoneme sets [24 English consonants in Cooke and Scharenborg, (2008) vs 14 consonants that occur both in German and English in this study]. Furthermore, Cooke and Scharenborg included VCV utterances with front and back stress, whereas in this study utterances produced with front stress and in various speaking styles were examined.

Sroka and Braida (2005) analyzed consonant confusions of human and automatic recognizers in speech-weighted noise with VCV utterances. In terms of the SNR, they observed a human–machine gap of 12 dB, which closely resembles the scores gathered in the study, when only consonant confusions for normal speaking style are considered (which results in a difference between human and machine recognition performance of 11.8 dB for Set *RES*). However, when vowel phonemes are included and scores are averaged over the complete test set, the gap in terms of the SNR is 15 dB (Set *RES*) and 13 dB (Set *DA*).

One of the main differences between HSR and ASR is the strategy that human listeners employ to detect speech components in a signal. For example, Miller and Licklider (1950) performed an experiment with interrupted (gated) speech and found that word recognition scores are only slightly degraded when the interruptions occur at modulation frequencies between 10 and 100 Hz. The authors assumed that a high intelligibility could be obtained as long as listeners get a glimpse at each phoneme of the presented word. Such a technique that would effectively allow for “listening in the dips” is not incorporated in current standard ASR systems. However, in Cooke (2005) a model based on spectro-temporal patches above the noise floor (or glimpses) is proposed, which was found to produce accurate predictions of human phoneme recognition rates. In the light of the recognition performance gap between humans and machines, an inclusion of such detection in ASR might therefore be worthwhile.

B. Effect of resynthesis

Experiments with resynthesized speech showed that the human–machine gap narrows when the information provided to human listeners is limited to the information contained in cepstral features. Compared to HSR with resynthesized signals, relative ASR error rates are only 29% higher in identical SNR conditions for Set *RES* (Set *DA:* 20%).

Earlier studies have shown that the information contained in ASR features based on short-term spectra 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. Leonard (1984) measured a 99.9% HSR recognition rate when presenting digits that were resynthesized from features coding the spectral envelope (i.e., linear prediction coefficients). Demuynck (2004) reported that clean speech which was resynthesized from MFCC features remains perfectly understandable.

However, the presented measurements based on noisy speech clearly show that high rates at high SNRs are due to the ceiling effect and that during the calculation of MFCCs a significant amount of useful information is removed, confirming the qualitative results from (Peters *et al.*, 1999). Since ASR results were obtained for several SNRs, the performance drop that is encountered in noisy environments (with the audible information being limited to cepstral information) can be expressed in terms of the SNR. For the phoneme recognition task based on the German phoneme set employed here, it can be estimated to be 10 dB. Compared to the gap of 15 dB that was found for the overall gap, it therefore appears that the major part of the human–machine gap can be attributed to an imperfect representation of speech used in standard ASR systems.

On the other hand, even when only using resynthesized signals from Set *RES*, humans outperformed the ASR system, since comparable ASR scores were obtained only when the SNR was increased by 5 dB (i.e., from 3.8 to 8.8 dB). Apparently listeners are better able to identify phonemes even when only resynthesized features are used, which hints at the suboptimal use of available cues by the hidden Markov system. This comparatively smaller contribution of the backend is consistent with the findings of Jürgens *et al.* (2009), who used an auditory model as front-end for ASR and compared the case of perfect a priori knowledge of the word to be recognized to a realistic ASR task, i.e., supplying imperfect a priori knowledge of the word to be recognized. While in the first case a near-to-perfect prediction of human recognition scores was possible, in the latter case a gap between observed and modeled speech recognition of approximately 13 dB was observed.

Following the physiological interpretation, the (stimulus-driven) bottom–up processing that corresponds to the feature extraction stage (as described in Sec. I) contributes the main part to the differences between the recognition performance of man and machine. When ASR techniques are employed to model human speech perception [as it has been done in (Cooke, 2005)], it might therefore be worthwhile to replace this processing scheme with the output of an auditory model.

Apart from the information loss due to feature calculation, other factors might contribute to the degraded speech intelligibility of resynthesized speech. The algorithm that was employed might not optimally reconstruct the time signals, i.e., not all information from the ASR features is perfectly made audible for the listeners. We tried to cover this problem as good as possible by performing pilot experiments (cf. Sec. II B 2) with various excitation signals; however, there might still be room for improvement by optimizing, e.g., the pulse form of the excitation signals. Training effects might also play a role in HSR since a fixed fundamental frequency was used for the excitation signal; this resulted in utterances that sounded artificial and unfamiliar to the listeners. Training sessions were performed before each measurement to familiarize the listening subjects with the stimuli, thereby limiting the influence of such training effects (Sec. III A 3).

The analysis based on transmitted information showed that resynthesized signals result in comparatively low scores for vowel recognition and vowel-associated features (Fig. 2). It therefore appears that the process of feature calculation and resynthesis affects human vowel recognition stronger than consonant recognition. 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. The fact that for resynthesized signals a higher SNR is required to achieve the same performance level (compared to unaltered signals) may also be caused by discarding the phase information which is in accordance with other works: Schlüter and Ney (2001) reported improvements of ASR scores when exploiting the phase information, and Peters *et al.* (1999) measured a degraded HSR recognition performance when the audible information was limited to the power spectrum.

C. Effect of intrinsic variations

Changes in speaking style, rate, and effort were found to degrade HSR (with an average relative degradation of 23%) and ASR (47% relative degradation in average at a SNR of 8.8 dB). The robustness of the ASR system against extrinsic and intrinsic variations can also be expressed in terms of the equivalent change in SNR. In ASR, the effect of intrinsic variations (Set *RES*) had the same effect as an increase of 5 dB of the SNR level. This estimation is based on two observations: The ASR accuracy increases almost linearly with approximately 2% per dB for SNRs from −5 to 15 dB; second, the average accuracy for normal speech is approximately 10% higher than speech with changed speaking rate, effort, and style.

Dialect and accent were shown to significantly affect HSR and ASR recognition scores. Furthermore, an analysis of variance (Table VI) showed that speaking rate, effort, and style, as well as the choice of speaker, contribute considerably to the variance of recognition scores in HSR and ASR. Intrinsic variations also have a significant effect on resynthesized speech. In this condition, however, the choice of speaker seems to have a more dominant effect than speaking rate, effort, and style. This might result from the elimination of speaker-specific, non-redundant cues (e.g., fine phonetic detail) that are removed during feature calculation and resynthesis. The remaining cues may not be sufficient for human listeners to adjust to speaker-specific changes. In case of ASR, the contribution of changes in rate, effort, and style are more important than in HSR with resynthesized speech, which is consistent with the high sensitivity of ASR against such variations (as described above).

1. Interaction between intrinsic variations and AFs

Sroka and Braida (2005) reported that the voicing feature was suboptimally recognized by a MFCC-based ASR system when testing phoneme recognition in speech-shaped noise. This finding is confirmed by the presented measurements: Compared to human performance, the voicing feature in ASR is degraded by 32% (averaged over the SNRs shown in Fig. 2) when utterances are spoken normally, while manner and place are degraded by 18% and 25%, respectively. The degraded recognition of voicing was expected, since the spectral fine structure is discarded during the calculation of MFCC features, which increases the confusions of voiced phoneme with their unvoiced counterparts. Nevertheless, these cues are also unavailable for a standard ASR system; the results suggest that a fine structure should be considered for improving ASR systems.

However, when different speaking styles are analyzed, large variations of the transmitted information of AFs were observed in this study. This was also found for HSR with resynthesized features, which indicates that MFCCs do not capture all the information required to deal with changes in speaking rate, effort, and style. These findings suggest the incorporation of some aspects of spectral fine structure associated with the recognition of voicing in order to overcome these deficiencies by changing the way in which spectral information is converted into speech cues. The use of such fine-grained cues seems especially important when the speech contains changes in speaking effort and rate.

Increased speaking effort (loud) produced the best consonant recognition for original signals in HSR. In spite of the higher SNR used for resynthesized signals, the T_r scores for resynthesized speech do not exceed those obtained with original utterances. In ASR, loudly spoken utterances result in the *lowest* consonant recognition. Similar to the observations for the voicing feature, it seems that human listeners rely on cues contained in speech spoken with increased effort that are lost during feature calculation, which at least partially contributes to the overall low ASR performance for high speaking effort.

T_r scores of resynthesized and original signals were consistent for the dialects under consideration, i.e., both conditions are similarly affected for the majority of the AFs. This suggests that variations caused by the dialects under investigation are equally well encoded by standard ASR features.

It is an open question which cues are the most relevant for this observed large difference between HSR and ASR. Obviously, several AFs are affected (cf. Fig. 2), but the perceptual cues associated with these features (especially for the AFs "place" and "manner") are not easily accessible. Although the recordings of the OLLO database have been carried out in a quiet environment, the properties of loudly spoken utterances may be similar to the parameters that are changed in Lombard speech, i.e., speech that has been recorded in noisy surroundings, which results in a slight decrease of consonant duration, changes in spectral properties of fricatives, and maximum burst energy of plosives (Junqua, 1993). Another possible explanation is the level normalization for loud utterances, since—for unnormalized speech—human listeners may rely on level differences in order to identify the speaking effort and subsequently selecting the cues that are of special importance for this specific variability.

2. Relation of phoneme duration and recognition score

In an attempt to determine the acoustic–phonetic cues relevant in phoneme recognition, an analysis of vowel duration and its relation to recognition rates was performed. For consonants a clear dependency between the duration and the recognition score for a phoneme was not observed (neither in HSR nor in ASR). For vowel identification in HSR two vowel groups were identified for which the recognition score either increased or decreased with longer phoneme durations. This trend is consistent with studies investigating temporal cues in HSR: Hillenbrand *et al*. (1995) reported that phonemes from the group D1 exhibit a longer duration than phonemes from D2 when produced with normal speaking rate, and Phatak and Allen (2007) showed phoneme duration to be an important cue in phoneme recognition.

The separation between D1 and D2 observed in HSR suggests that human listeners rely on these temporal cues for vowel recognition in situations in which target phonemes with otherwise similar properties (e.g., formant frequencies) need to be distinguished. Temporal cues also seem to play a major role for differentiation between vowels in resynthesized speech, while the ASR system seems not to rely on these cues in high noise conditions. This indicates that human listeners employ different strategies to process the information from ASR features compared to a HMM, e.g., by temporal integration of the signal, or by recognizing the patches of the internal representation belonging to the acoustic object that should be recognized.

Techniques that explicitly account for the temporal dynamics of speech have been reported to lower error rates in many acoustical scenarios (Hermansky and Sharma, 1999). It remains to be seen if these techniques can help to decrease the degradations in the presence of altered speaking rate.

Bronkhorst *et al*. (1993) have shown that the recognition performance increases when meaningful CVCs are presented instead of nonsense CVCs. Such an increase is therefore expected when analyzing continuous conversational, meaningful speech instead of the logatomes employed in the current study. Therefore, the influence of the specific intrinsic variations investigated in this study on conversational speech has yet to be quantified. Variations in conversational speech are considerably larger than recordings under controlled situations, as speaking rate and effort are subject to frequent changes. However, experiments comparable with our approach would require a database with labeled phonemes and variabilities, which does not yet exist to our knowledge. For the creation of suitable databases, problems such as the ambiguous labeling of phonemes are further aggravated in the presence of strong variations in spoken language, as, e.g., Shriberg *et al*. (1984) have shown for the transcription of children speech.

V. CONCLUSIONS

Even for the relatively simple task of phoneme classification, the difference between human and ASR performance remains considerably large. The relative increase of errors for ASR systems is larger than 150% (assuming medium speech intelligibility). If the information contained in standard ASR features is made audible and presented to human listeners, the gap narrows, but the relative ASR error rates are still at least 20% higher.

The information loss caused by the calculation and resynthesis of MFCCs can be expressed in terms of the SNR: Comparable recognition results in HSR are obtained when the SNR is 10 dB higher for resynthesized signals compared to unaltered speech files. The average performance gap between human and automatic recognition in terms of the SNR was found to be approximately 15 dB. The experiments with resynthesized speech showed that a 10 dB increase of the SNR is required to compensate for the information loss that arises from the resynthesis. Therefore, the contribution of the feature extraction to the human–machine gap of 15 dB can be estimated to be 10 dB. Nevertheless, human listeners outperformed ASR systems even when the acoustic information was limited to the information that is supplied to conventional ASR systems. The ASR classification score reached the human performance level only when the SNR was increased by 5 dB, which can therefore be seen as an estimate for the contribution of the HMM to the human–machine gap.

Speech-intrinsic variations were shown to significantly affect both human and machine performance and increased word error rates by up to 120% (Fig. 2). The analysis based on AFs showed that for utterances with increased speaking effort and high speaking rate, the differentiation between voiced and unvoiced sounds was especially problematic in ASR. A way to cope with this deficiency may be to modify the scheme of purely spectral features (e.g., by introducing feature components that cover some aspects of spectral fine structure).

An analysis of the relation of phoneme duration and the phoneme recognition rate showed no consistent trends between duration and the classification of consonants. However, the recognition rates of vowels heavily depend on the duration of these phonemes. Both in HSR and ASR, two groups of vowels were identified that yielded either an improved or a deteriorated recognition with increased duration. While the errors in HSR were consistent over a wide range of SNRs, the ASR confusion patterns were less consistent. This inability of the ASR system to utilize duration cues in a similar way as in HSR suggests that temporal and spectro-temporal aspects of speech should be incorporated in ASR systems in a more appropriate way, which might be better suited to capture vowel transients.

VI. ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgesellschaft within the collaborative research centre “The active auditory system” (URL: http://www.uni-oldenburg.de/sfbtr31). The Oldenburg Logatome Speech Database has been developed as part of the EU project DIVINES (Diagnostic and intrinsic variabilities in natural speech, IST-2002-002034). We would like to thank Thorsten Wesker, Matthias Wächter, Alfred Mertins, Jörn Anemüller, and Kirsten Wagener for their support and contributions to this work. We also thank Kris Demuynck, who kindly supplied the resynthesis algorithm.

[1] The OLLO database, including a detailed description, wordlists, labeling files, technical specifications and calibration data [normalization coefficients and dB (SPL) values], is freely available for research in HSR and ASR. The uncompressed corpus is approximately 6.4 GB in size and contains a total of approximately 140 000 files corresponding to 60 h of speech. It can be downloaded from http://medi.uni-oldenburg.de/ollo (last viewed on June 20, 2010).

[2] http://www.hoertech.de (last viewed on June 20, 2010)

Barker, J., and Cooke, M. (**2007**). “Modelling speaker intelligibility in noise,” Speech Commun. **49**, 402–417.

Benzeguiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L, Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (**2007**). “Automatic speech recognition and speech variability: A review,” Speech Commun. **49**, 763–786.

Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (**1993**). “A model for context effects in speech recognition,” J. Acoust. Soc. Am. **93**, 499–509.

Carey, M. J., and Quang, T. P. (**2005**). “A speech similarity distance weighting for robust recognition,” in *Proceedings of Interspeech*, Lisbon, Portugal, pp. 1257–1260.

Cooke, M. (**2005**). “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am. **119**, 1562–1573.

Cooke, M., and Scharenborg, O. (**2008**). “The Interspeech 2008 consonant challenge,” in *Proceedings of Interspeech*, pp. 1781–1784.

Davis, S., and Mermelstein, P. (**1980**). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust. Speech. Signal Process. **28**, 357–366.

Demuynck, K., Garcia, O., and Van Compernolle, D. (**2004**). “Synthesizing speech from speech recognition parameters,” in *Proceedings of Interspeech*, pp. II-945–II-948.

Dreschler, W. A., Verschuure, H., Ludvigson, C, and Westermann, S. (**2001**). “ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment,” Int. J. Audiol. **40**, 148–157.

Gillick, L., and Cox, S. J. (**1989**). “Some statistical issues in the comparison of speech recognition algorithms,” in *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, United Kingdom, pp. 532–535.

Griffin, D., and Lim, J. (**1984**). “Signal estimation from modified short-time Fourier transform,” IEEE Trans. Acoust. Speech. Signal Process. **32**, 236–243.

Hawkins, S. (**2003**). “Roles and representations of systematic fine phonetic detail in speech understanding,” J. Phonetics **31**, 373–405.

Hermansky, H., and Sharma, S. (**1999**). “Temporal patterns (TRAPS) in ASR of noisy speech,” in *Proceedings of the 1999 International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, pp. 289–292.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). “Acoustic characteristics of American English vowels,” J. Acoust. Soc. Am. **97**, 3099–3111.

Junqua, J.-C. (**1993**). “The Lombard reflex and its role on human listeners and automatic speech recognizers,” J. Acoust. Soc. Am. **93**, 510–524.

Jürgens, T., Brand, T., and Kollmeier, B. (**2009**). “Predicting consonant recognition in quiet for listeners with normal hearing and hearing impairment using an auditory model (A),” J. Acoust. Soc. Am. **125**, 2533.

Kipp, A., Wesenick, M., and Schiel, F. (**1996**). “Automatic detection and segmentation of pronunciation variants in German speech corpora,” in *Proceedings of the 1996 International Conference on Spoken Language Processing*, Philadelphia, Pennsylvania, pp. 106–109.

Kohler, K. (**1995**). *Einführung in die Phonetik des Deutschen (Introduction to German phonetics)* (Erich Schmidt Verlag, Berlin), pp. 1–249.

Krause, J. C., and Braida, L. D. (**1995**). “The effects of speaking rate on the intelligibility of speech for various speaking modes (A),” J. Acoust. Soc. Am. **98**, 2982.

Krause, J. C., and Braida, L. D. (**2003**). “Acoustic properties of naturally produced clear speech at normal speaking rates,” J. Acoust. Soc. Am. **115**, 362–378.

Leonard, R. (**1984**). “A database for speaker-independent digit recognition,” in *Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing*, Vol. IX, pp. 328–331.

Lippmann, R. (**1997**). “Speech recognition by machines and humans,” Speech Commun. **22**, 1–15.

Meyer, B., and Wesker, T. (**2006**). “A human-machine comparison in speech recognition based on a logatome corpus,” in *Workshop on Speech-Intrinsic Variation*, pp. 95–100.

Miller, G. A., and Licklider, J. (**1950**). “The intelligibility of interrupted speech,” J. Acoust. Soc. Am. **22**, 167–173.

Miller, G. A., and Nicely, P. E. (**1955**). “An analysis of perceptual confusions among some English consonants,” J. Acoust. Soc. Am. **27**, 338–352.

Peters, S., Stubley, P., and Valin, J. (**1999**). “On the limits of speech recognition in noise,” in *Proceedings of the 1999 International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, pp. 365–368.

Phatak, S. A., and Allen, J. B. (**2007**). “Consonant and vowel confusions in speech-weighted noise,” J. Acoust. Soc. Am. **121**, 2312–2326.

Scharenborg, O. (**2007**). “Reaching over the gap: A review of efforts to link human and automatic speech recognition research,” Speech Commun. **49**, 336–347.

Schlüter, R., and Ney, H. (**2001**). “Using phase spectrum information for improved speech recognition performance,” in *Proceedings of the 2001 International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, pp. 133–136.

Shen, W., Olive, J., and Jones, D. (**2008**). “Two protocols comparing human and machine phonetic recognition performance in conversational speech,” in *Proceedings of Interspeech*, pp. 1630–1633.

Shriberg, L. D., Kwiatkowski, J., and Hoffmann, K. (**1984**). “A procedure for phonetic transcription by consensus,” J. Speech Lang. Hear. Res. **27**, 456–465.

Sroka, J. J., and Braida, L. D. (**2005**). “Human and machine consonant recognition,” Speech Commun. **45**, 401–423.

Stern, R., Acero, A., Liu, F. H., and Ohshima, Y. (**1996**). “Signal processing for robust speech recognition,” in *Automatic Speech and Speaker Recognition*, edited by C.-H. Lee, F. K. Soong, and K. K. Paliwal (Springer, Berlin), pp. 357–384.

Wesker, T., Meyer, B., Wagener, K., Anemueller, J., Mertins, A., and Kollmeier, B. (**2005**). “Oldenburg Logatome Speech Corpus (OLLO) for speech recognition experiments with humans and machines,” in *Proceedings of Interspeech*, pp. 1273–1276.