Human phoneme recognition depending on speech-intrinsic variability[a)]

Bernd T. Meyer,[b)] Tim Jürgens, Thorsten Wesker, Thomas Brand, and Birger Kollmeier
Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, D-26111 Oldenburg, Germany

(Received 29 December 2008; revised 2 September 2010; accepted 3 September 2010)

The influence of different sources of speech-intrinsic variation (speaking rate, effort, style and dialect or accent) on human speech perception was investigated. In listening experiments with 16 listeners, confusions of consonant-vowel-consonant (CVC) and vowel-consonant-vowel (VCV) sounds in speech-weighted noise were analyzed. Experiments were based on the OLLO logatome speech database, which was designed for a man-machine comparison. It contains utterances spoken by 50 speakers from five dialect/accent regions and covers several intrinsic variations. By comparing results depending on intrinsic and extrinsic variations (i.e., different levels of masking noise), the degradation induced by variabilities can be expressed in terms of the SNR. The spectral level distance between the respective speech segment and the long-term spectrum of the masking noise was found to be a good predictor for recognition rates, while phoneme confusions were influenced by the distance to spectrally close phonemes. An analysis based on transmitted information of articulatory features showed that voicing and manner of articulation are comparatively robust cues in the presence of intrinsic variations, whereas the coding of place is more degraded. The database and detailed results have been made available for comparisons between human speech recognition (HSR) and automatic speech recognizers (ASR).

© *2010 Acoustical Society of America.* [DOI: 10.1121/1.3493450]

PACS number(s): 43.71.Es, 43.71.Gv, 43.72.Ne [DOS] Pages: 3126–3141

I. INTRODUCTION

Normal human listeners exhibit an excellent performance in speech recognition despite the immense variations present in spoken language. This holds even if different speakers have to be understood, i.e., human listeners can compensate for a variety of speaking rates, different regional accents and different vocal effort of the received speech.

The robustness to these underlying speech intrinsic variabilities is a major achievement of human speech recognition (HSR) that is not well understood yet. Many sources of variation in spoken language have been observed and well documented in several studies, such as, for example, the gender and age of the talker (male versus female speaker versus children speech (Hazan and Markham, 2004), the effect of certain speaking styles [such as, e.g., speaking clearly to achieve a higher intelligibility (Krause and Braida, 2004)], and the influence of dialect and accent on speech intelligibility (Li, 2003). Other factors that may influence speaking rate and effort are, e.g., emotion, stress, fatigue, and health condition. These sources of variation are not independent, an example being the influence of speaking rate on pronunciation that arises from deletions, insertions, and coarticulation (Fosler-Lussier and Morgan, 1999). Despite the large number of studies dealing with variations in speech, it is still unclear how the auditory system manages to produce percepts that are largely invariant to such changes in speech.

While the robustness of automatic speech recognition (ASR) against extrinsic variability (caused by changes of the transmission channel or by additive noise) has been studied in detail in the past (Hermansky and Morgan, 1994; Stern *et al.*, 1996; Tchorz and Kollmeier, 1999; Cooke *et al.*, 2001), it is far less understood in which way ASR also suffers from a lack of robustness toward "intrinsic" variation of speech (caused by factors such as the choice of speaker, gender, speech rate, vocal effort, regional accents, and speaking style). Various methods that increase the robustness of ASR toward such variation of speech have successfully been used for several years (an example being techniques that compensate for the shift of formant frequencies caused by variations of vocal-tract length). Recently, however, much of the research is devoted to a larger number of sources of variability, with the aim of understanding the influence of speech-intrinsic variation on ASR, and to build feature extraction or classification methods invariant to this variation (Fissore *et al.*, 2007). Similarly, several researchers focused on the comparison of the recognition performance of HSR and ASR (Lippmann, 1997; Sroka and Braida, 2005; ten Bosch and Kirchhoff, 2007; Cooke and Scharenborg, 2008). Meyer *et al.* (2007) used a phoneme recognition task to compare HSR and ASR and observed comparable overall results when the signal-to-noise ratio was 15 dB higher for the ASR system. In that study, the larger part of the man-machine gap was attributed to the feature extraction stage. However, the 15 dB gap is only a rough estimate that is highly dependent on the exact type of experiment to be compared across men and

[a)] Parts of this work were presented at the Eighth Annual Conference of the International Speech Communication Association (Interspeech 2007, Antwerp).

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

TABLE I. Properties of the OLLO speech database.

Number of speakers	50 (25 male, 25 female)
Number of different VCVs	70 (five outer vowels (/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/) combined with 14 central consonants (/b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /s/, /ʃ/, /t/, /v/, ts/))
Number of different VCVs	80 (eight outer consonants (/b/, /d/, /f/, /g/, /k/, /p/, /s/, /f/) combined with 10 central vowels (/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/, /a:/, /e/, /i/, /o/, /u/))
Number of different logatomes	150
Number of speaking styles	5+reference condition ‘normal’ (fast, slow, loud, soft, question)
Number of dialects/accents	4+reference condition ‘no dialect’ (East Frisian, Bavarian, East Phalian, French)
Utterances per speaker	2700 (150 logatomes × 3 repetitions × 6 speaking styles)
Total number of logatomes	133.403
Utterances labeled as containing unwanted sounds	1820
Number of utterances per dialect/accent	~27 000
Number of utterances per variability	~22 200
Number of utterances per central consonant	~4450
Number of utterances per central vowel	~7100

machines. Such comparisons highlight the deficiencies of current automatic recognizers in the presence of extrinsic and intrinsic variation of spoken language.

Since understanding the principles of HSR may help to improve the performance of ASR (Allen, 1994), it is therefore desirable to study the influence of speech variability on HSR as a baseline for making ASR more robust against this type of variation. For example, it was shown that error rates increase when speaking rates deviate from the normal (i.e., average) speaking rate. Siegler and Stern (1995) reported an increase of ASR error rates by a factor of three when the rate of speech deviated more than two standard deviations from the average rate. The effect of conversational speech was investigated by Weintraub *et al.* (1996) who found that error rates doubled when conversational speech is compared to a read, clearly uttered version of the same speech material. However, it is often difficult to compare findings from different HSR and ASR studies due to the existing variability across speakers in the available speech databases and the lack of appropriate speech corpora that are suitable both for HSR and ASR experiments while providing the possibility to study the effect of speech-intrinsic variation.

In this study, we therefore perform HSR experiments to assess the impact of several sources of intrinsic variability by using a speech database that is suitable for HSR and ASR experiments and that contains systematically varied sources of variability. This database (the ‘Oldenburg Logatome (OLLO) Corpus’) consists of CVC and VCV utterances, which we refer to as logatomes (Table I). The logatomes are composed according to phonetic and phonotactic rules, and—in most cases—have no semantic meaning for German and English listeners. The choice of phonemes included in the database was based on earlier results from HSR and ASR experiments. In a few cases, the combination of these phonemes resulted in meaningful CVCs and VCVs; nevertheless, these logatomes were used for the experiments in order to avoid the introduction of a bias due to the selection of phonemes. By using simple nonsense phoneme combinations, the focus is laid on a basic recognition task that does not rely on high-level lexical knowledge. Such recognition can primarily be considered as a bottom-up sensory one-out-of-N discrimination task in HSR that requires no prior knowledge of the language structure and a low cognitive load imposed on the listeners when performing the task. In ASR the recognition task requires templates or word models primarily on the acoustical feature layer without a suprasegmental or language model to be fitted to the speech data. Hence, the OLLO database can be used as a reference for HSR research as well as for comparing ASR experiments using the same speech elements. Moreover, the influence of speech variability on both types of experiments can easily be studied. The principles underlying the database construction and its recording will be discussed. It is assumed that the results obtained in phoneme recognition tests may also be of value for other tasks such as large vocabulary ASR, even when the results may not be directly transferable since the use of context knowledge has to be taken into account (Bronkhorst *et al.*, 1993).

The primary aim of the current paper is to establish the baseline for HSR experiments with the OLLO that can be utilized in future work for comparison with ASR. To do so, the influence of speaking style (i. e., fast, slow, soft, loud) as well as speaker-specific factors (gender and dialect region) on HSR is studied with a total number of 16 listeners and 120 h of listening experiments. Speakers originated from various dialect regions in Germany and from the French-speaking part of Belgium, which enabled an analysis of the effect of dialect and accent. Since all phonemes in the database occur both in the German and English languages, the utterances may also be of useful for listening tests with English listeners. The results presented in this study were obtained with German listeners. Even though some differences in average recognition rates from the mentioned variabilities are expected (especially when the experiment is performed in noise, which is necessary to avoid any ceiling effect), it is unclear if these differences are due to the deterioration of specific speech features or due to a general, unsystematic decrease in intelligibility. For this reason, a speech transmission analysis (Miller and Nicely, 1955; Wang and Bilger, 1973) should be performed that studies the transmission of acoustic speech features (such as, e. g., average spectrum of the phonemes to be recognized or articulatory features) as a function of underlying speech variabilities. In order to cancel out the individual influence of each individual listener, such an analysis only makes sense if an appropriate amount of data is available that can be averaged across listeners. Hence, the number of subjects was selected to be sufficiently high to derive valid conclusions for these aspects of HSR.

This paper is structured as follows: In Section II, a detailed description of the Oldenburg Logatome speech database is presented. The measurement setup, parameters for the listening tests and outcome measures for data analysis are described in Section III. Overall results and effects of variabilities on information transmission are presented in Section IV. Section V and Section VI contain the discussion of results, a summary and the conclusions.

II. CREATION AND DEVELOPMENT OF THE OLLO CORPUS

This section describes the design choices for the creation and the development of the Oldenburg Logatome Corpus. For the listening experiments performed in this study, several subsets of the database were selected, which are described in Section III A.

A. Choice of phonemes and speech stimuli

The corpus used for this study should contain speech with labeled, speech-intrinsic variabilities. The experiments aim at the simple task of phoneme recognition without the possibility to exploit context knowledge. An analysis of coarticulation effects and easy determination of phoneme recognition rates are further desirable properties. Short combinations of phonemes satisfy all of these pre-requisites. We chose combinations of vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) with identical outer phonemes for the database. The standard recognition task for those nonsense utterances or logatomes is to identify the middle phoneme, which limits the number of response alternatives and allows for an easy realization of HSR tests. Since the OLLO corpus should be suitable for a comparison of speech recognition by human listeners and automatic speech recognizers, the choice of phonemes was based on HSR and ASR recognition experiments. Phonemes that are critical in either human or automatic recognition of speech were selected, so that significant differences in recognition rates may already be obtained with smaller test sets.

1. Critical phonemes in human and automatic speech recognition

The results of monosyllabic and bisyllabic rhyme tests with normal-hearing listeners were analyzed to determine the phonemes that are most often confused by human listeners in English or German (Dubno and Levitt, 1981; Gelfand *et al.*, 1985; Müller, 1992; Kliem, 1993). The results suggest that eleven consonant phonemes (/b, d, f, g, k, l, p, r, s, v, ts/) and seven vowel phonemes (/æ, ε, ɪ, i, ʊ, u, y/) should be taken into account.

In order to determine the critical phonemes in ASR, phoneme confusions from a recognition experiment were analyzed: Eight phonemes (/s, ʃ, l, k, m, n, p, t/) were selected for the corpus because they produced high error rates, appear in both the German and English languages, and are often present in phoneme confusions.

2. Final set of phonemes

The final number of phonemes to be considered was limited by the required time to record all necessary items with a single speaker. Since the standard recognition task for the OLLO database is to identify the middle phoneme, not all possible combinations of consonant and vowel phonemes were taken into account. The final phoneme set for VCVs consists of five vowel phonemes (/a, ε, ɪ, ɔ, Ʊ/) and 14 consonant phonemes (/b, d, f, g, k, l, m, n, p, s, ʃ, t, v, ts/). The set for CVCs contains one of ten vowels (/a, ε, ɪ, ɔ, Ʊ, a:, e, i, o, u/) and one of eight consonants /b, d, f, g, k, p, s, t/). A combination of these phonemes results in a total of 150 different logatomes (70 VCVs, 80 CVCs). The vowels are different with respect to height, backness and roundedness (i.e., their constituent features in the cardinal vowel system [MacArthur, 1992)], with the exception of /a/ and /a:/, which differ only by a suprasegmental indicating different phoneme durations.

B. Sources of variability and speaker selection

The choice of different sources of variability was based on ASR experiments with annotated test corpora that compared the performance of automatic recognizers with these specific variations present or not. The sources under consideration included speaker's gender, age and dialect, speaking style/effort (which also relates to pitch), rate of speech, and breathing noise. The largest impact on performance was observed for the speaking rate (fast vs. slow), speaking style (affirmation vs. question), speaking effort (loud vs. soft), and dialect/accent. The latter was integrated in the database by including logatomes of dialect speakers from different regions of Germany and from the French-speaking part of Belgium. Ten speakers originating from the northern part of Germany (Oldenburg near Bremen and Hanover) were recorded. The spoken language in this region is usually considered as standard German (Kohler, 1995). We will refer to this category as ‘no dialect’ (ND). Furthermore, speakers were recorded who originate from the Northwestern part of Germany and commonly speak the East Frisian Lower Saxon dialect (abbreviated EF) rather than standard German. Other dialects were included by recording speakers from East Phalia (EP) near Magdeburg, and from Bavarian places near Munich (BV). The French-speaking participants were recorded in Mons (Belgium). They did not speak German as a second language and usually produce (and perceive) the same phonemes differently from the German population, e.g., the French voiceless stop is more similar to the German voiced stop than to the German voiceless one. This group was included in order to be able to test the influence of different phoneme boundaries on HSR and ASR. Five female and five male speakers from each region were recorded, resulting in a total of 50 speakers. The age of speakers varied between 18 and 65 years. Each logatome was recorded in a ‘neutral/clear’ speaking style as a reference. In addition, one of the five selected categories (i.e., fast and slow speaking rate, loud and soft speaking style, and condition ‘question’, which refers to rising pitch) was alterered for each of the subsequent recordings. Note that combinations of these sources of variability were not recorded (i.e., utterances with high speaking effort and high speaking rate were not recorded, for instance).

To provide a broad test and training basis for ASR experiments and to enable an analysis of intra-individual differences, each logatome was recorded three times, which resulted in $150 \times (5+1) \times 3 = 2700$ logatomes per speaker. Additionally, for German speakers 72 German words that are part of the monosyllabic rhyme test (Kollmeier and Wallenberg, 1989) and 20 German sentences part of the Goettingen sentence test (Kollmeier *et al.*, 1997) were included. Participants from Belgium recorded 20 French sentences. The sentences are phonetically balanced and can be used to perform an adaptation of model parameters in automatic recognizers to individual speakers.

C. Recording setup

1. Technical equipment

All utterances were recorded in sound-insulated audiometry rooms (reverberation time: approximately 0.25 s) with a studio-quality condenser microphone (AKG C1000 S) placed approx. 30 cm from the speaker. Recordings were carried out using a RME QuadMic microphone pre-amplifier and an RME Hammerfall AD converter connected to a standard notebook. The software for the presentation of logatome transcriptions and for recording was based on Matlab (The MathWorks) and SoundMex (HoerTech GmbH). The original sampling frequency was 44.1 kHz at 32 bit resolution, which was reduced to 16 kHz and 16 bit during post processing.

2. Recording conditions

Since the database was intended to contain speech from phonetically naïve speakers, a transcription of the desired logatome and speaking style was created by a phonetician and presented to speakers on a computer screen (an example being ‘Please speak ascha loudly’, where ‘ascha’ is intepreted as /a ʃ a/ by German speakers). An adjustment of transcriptions was carried out for recordings of French speakers as well. Special attention was paid to the transcription and pronunciation of the near-closed phonemes /ɪ/ and /ʊ/, which are absent from French. Typographic accents and duration markers were used for the transcription. However, control samples showed that a considerable part of vowels embedded in CVCs is nevertheless categorized as closed phonemes /i/ and /u/ by linguists and the majority of German listeners. This is due to the fact that non-native speakers replace unfamiliar phonemes in the target language, which is absent in their native laguage phoneme inventory, with the sound considered as the closest in their native language phoneme inventory (Flege *et al.*, 2003). This replacement is likely to increase errors for speakers with accent.

Randomized sequences of 150 logatomes with the same variability were recorded. After each run, a different variability was randomly chosen for the next sequence. Speakers were supervised during the recordings and periodically reminded to speak in the desired manner. All VCV stimuli were produced with front stress. During training sessions, speakers could familiarize themselves with the recording software, which included proceeding to the next item by pressing a key on a keyboard and the option for re-recording of utterances that were contaminated with unwanted sounds or were not judged by the speaker or the supervisor to be uttered in the appropriate way. Speakers were advised to speak in a natural manner; the realization of variabilities was checked and corrected if necessary. Some of the logatomes that contain a short vowel embedded in plosives (e.g., /p a p/) cannot be spoken slowly. Speakers were asked to articulate the logatome with normal speaking rate when the desired variability would conflict with the pronunciation. Participants were encouraged to take regular breaks to avoid mispronunciation due to inattentiveness. The average duration of the whole recording procedure was 3.5 h per speaker.

D. Postprocessing of recorded material

A quality check of the recordings was carried out using a semi-automatic software written in Matlab that relied on a simple energy criterion to detect incomplete utterances or recordings with an audible keystroke. Unwanted sounds coinciding with the silence before or after the utterance were manually removed from the signal. 1597 signals (which corresponds to 1.2% of the total number of utterances) that were incomplete or had background noise in the speech signal were removed from the database. Another 1820 utterances (or 1.4%) were labeled as containing a quiet, unwanted sound, which is audible in silence, but not in the presence of noise. Effects caused by these sounds are assumed to be negligible for the measurements presented in this work, as subsets of OLLO were chosen for listening tests, and unsuitable utterances were removed from those sets. The silence at the beginning and at the end of each recording was limited to 500 ms. Signals were then normalized to 99% amplitude and stored with 16 bit resolution. They were low-pass filtered with an 8 kHz cutoff frequency and sampled down to 16 kHz.

E. Phonetic labeling

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 procedure is similar to forced alignment approaches based on hidden Markov models (HMMs). However, in contrast to standard forced alignment, it has the ability to take into account pronunciation variations typical to a given language by computing a statistically weighted graph of all likely pronunciation variants. For details, the reader is referred to (Kipp *et al.*, 1996).

All 150 logatomes were transcribed into the SAM phonetic alphabet (SAMPA) and the transcription was used as input for the time-labeling procedure. The MAUS labeling tool was applied to the data in 'full mode', i.e., taking into account pronunciation variations of the German language, and in addition the same software was applied in 'align-only' mode where HMM forced alignment is performed, but pronunciation variants are not considered.

In about 4.7% of the logatomes, the MAUS method's result deviated from the forced alignment result. Most of these differences (75%) can be accounted for by negligible shifts in phoneme boundary positions. The remaining quarter of the utterances with deviating boundaries had a pronunciation variant identified by MAUS. Most of such variations corresponded to shifts from short vowel forms (e.g., [a]) to the longer form (e.g., [a:]), which are plausible variations of the orthographic transcript presented to the speakers. The relative rarity of such variations indicates that in the vast majority of utterances the chosen orthographic transcript was pronounced in the way intended by the experimenters.

F. Availability of speech material and test results

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 approx. 6.4 Gbyte 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.

TABLE II. Subsets of the OLLO database used for human listening tests. The sets are used to analyze the influence of variabilities such as speaking rate, effort and style (Set *RES*), dialect or accent (Set *DA*) and SNR (Set *SNR*). Each set contains at least 150 different logatomes with 24 central phonemes. For sets with more than one speaker, the gender is equally distributed.

Aim of experiment: Analysis of the effect of speaking rate, effort and style (Set RES):	
Masking noise	Stationary, speech-shaped noise (−6.2 dB SNR)
HSR test set	Set *RES*: CVC and VCV utterances with two speaking rates (fast/slow), effort (loud/soft) and style ('question'/normal), Four talkers (2M, 2F) 3600 utterances (150 logatomes × 4 speakers × 6 speaking styles)
HSR listening subjects	Six normal hearing subjects (3M, 3F)
Aim of experiment: Analysis of the effect of dialect and accent (Set DA):	
Masking noise	Stationary, speech-shaped noise (−6.2 dB SNR)
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)
Aim of experiment: Analysis of the effect of signal-to-noise ratio (Set SNR):	
Masking noise	Stationary, speech-shaped noise (SNRs: −20, −15, −10, −5, 0 dB)
HSR test set	Set *SNR*: CVC and VCV utterances (normal speaking style, no dialect), one male talker, 750 utterances (150 logatomes × 5 SNRs)
HSR listening subjects	Ten normal hearing subjects (7M, 3F)

III. METHODS

A. Test sets and presented stimuli

Utterances from the OLLO databases were selected to analyze the effects of speaking style and effort, dialect and accent, and SNR. These selections are referred to as sets, and their properties have been summarized in Table II. The names were chosen according to the varied parameters in that set: Set *RES* is used to analyze the influence of speaking rate, effort, and style. It contains data in different speaking styles produced by four talkers (two male, two female) without regional dialect (ND=no dialect). Set *DA* is used to study the effect of dialect and accent, and contains utterances from two speakers (one male, one female) from each of the five dialect/accent regions with normal speaking style. From the 50 speakers in the database, those speakers were chosen as being representative for the corpus that produced recognition rates for a standard ASR task, which were closest to the average recognition rate. The experimental setup for this ASR test is described in Appendix B. The effect of a stationary, additive noise is investigated with Set *SNR* that contains data from a single male speaker. The data obtained with this

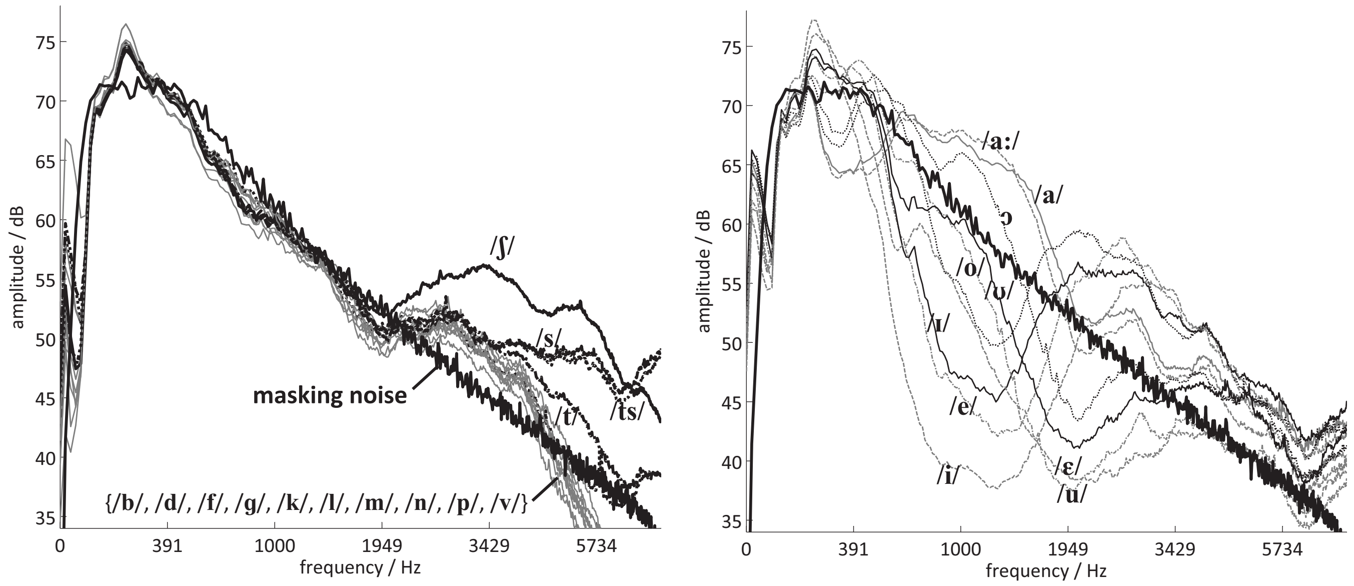

FIG. 1. Average DFT power spectrum of stationary masking noise signal (thick black line) and long-term spectra of OLLO utterances. Individual long-term spectra for central consonant and vowel phonemes are shown in the left and right panel, respectively. Mel-scaling with labels in Hz has been chosen for the frequency axis.

set can be used to compare the effects of intrinsic variations (quantified with Sets *RES* and *DA*) and additive noise.

When presented without masking noise, human listeners achieve very high phoneme recognition scores. For example, Meyer *et al.* (2006) measured scores > 99.5% for normally spoken VCVs and CVCs without accent or dialect. For accented, clean speech, the lowest recognition rate was found to be 95%. This high performance prevents a valid analysis of phoneme confusions, because differences at very high error rates often are outside the range of reliably observable differences (ceiling effect). Hence, the OLLO utterances were presented in noise. A stationary noise signal with speech-like frequency characteristics was added to the VCV and CVC utterances (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 have a close resemblance to real-life communication without clear modulation, equivalent to a situation with loud cocktail party noise. The original ICRA1 noise was downsampled from 44.1 kHz to 16 kHz using the Matlab resample function. The average DFT power spectrum of the resampled noise signal is shown in Fig. 1.

In order to identify the SNR at which recognition rates in the range of 70% to 80% are obtained, pilot measurements with a single listener and a small test set were performed. Based on these measurements, a fixed SNR of −6.2 dB was chosen for Sets *RES* and *DA*. For Set *SNR*, the utterances of one speaker (no dialect) and normal speaking style were used to analyze the dependency of recognition performance and noise. Speech-weighted noise at signal-to-noise ratios ranging from −20 dB to 0 dB was added to the logatomes. A summary of the Sets *RES*, *DA* and *SNR* is listed in Table II. The SNR was calculated based on the rms values of individual utterances. Due to variations of the duration of silence before and after each logatome for Set *DA*, a different SNR criterion was chosen for Set *DA* than for Set *RES*, which is described in Appendix C.

Figure 1 shows the long-term spectra of logatomes with different central phonemes. The long-term spectra were obtained by calculating the spectrum of each utterance, performing an rms-normalization and smoothing of each spectrum, and averaging over all spectra with identical central phonemes. The mean spectra were normalized to have the same rms level before plotting. By using this calculation scheme, the spectral properties of both vowels and consonants are represented in the long term spectrum. However, since for each central vowel the type and number of outer consonants is the same, the effects of outer phonemes are expected to average out.

B. Measurement setup and listeners

Sixteen German, normal-hearing listeners (10M, 6F) without regional dialect (cf. Section II F) participated in the HSR tests. From those sixteen subjects, six listeners (three male, three female) participated in the measurements with Set *RES*. Of those six, five listeners (three male, two female) also participated in the measurements with Set *DA*. Ten other listeners (7 male, 3 female) were chosen for Set *SNR*. The listeners were between 18 and 38 years old. Their hearing threshold for pure tones in standard audiometry did not exceed +20 dB at more than one data point and +10 dB at more than two data points in the pure tone audiogram. Randomized sequences of logatomes were presented in a soundproof booth and via audiological headphones (Sennheiser HDA200) after an online free-field equalization was performed. Feedback or the possibility to replay the logatome was not given during the test procedure. After a training phase, listeners were presented a sequence of logatomes at a level of 70 dB SPL, which was the preferred level of most

TABLE III. Phonetic features of eleven consonants. The articulatory feature ‘voicing’ can assume two feature values (voiced and unvoiced). For manner of articulation, consonants are categorized as stop, fricative or nasal. Possible values for place of articulation are anterior, medial, and posterior.

Consonant	p	t	k	b	d	g	s	f	v	n	m
Voicing	u	u	u	v	v	v	u	u	v	v	v
Manner	s	s	s	s	s	s	f	f	f	n	n
Place	a	m	p	a	m	p	a	m	p	m	a

listeners. For each presentation, a logatome had to be selected from a randomized list of CVCs or VCVs with the same outer phoneme and different middle phonemes, which triggered the presentation of the next random listening item after a short pause. The number of choice alternatives was either 10 (corresponding to the 10 central vowels used in CVCs) or 14 (since 14 different central phonemes are used for the VCVs). Carrier phrases were not used. A computer mouse was used as input device. In order to avoid errors due to inattentiveness, listeners were encouraged to take regular breaks. The total measurement time for each listener varied between 6 and 9 h, 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 each listener.

C. Data analysis

1. Articulatory features

It was analyzed whether the particular information associated with articulatory features (voicing, place and manner of articulation) was transmitted across the various speaking styles and dialects under investigation (e.g., if the voicing information is equally transmitted when the speaking rate changes). Miller and Nicely (1955) proposed the analysis of such articulatory features (AFs) with the aim of identifying the sources in consonant confusions. Recently, AFs have gained much interest in the field of automatic speech processing: For instance, AFs have been used to improve the recognition of phones in a language-independent ASR system (Siniscalchi *et al.*, 2008), and to increase the performance of multi-lingual annotation tools (Chang *et al.*, 2005). Since the study by Miller and Nicely, the analysis of AFs has become a standard way of summarizing phoneme confusions: Many studies investigating phoneme recognition used AF confusions as an analysis tool with the aim of indentifying specific sources of phoneme errors that relate to articulatory gestures of the vocal tract (e.g., Dubno and Levitt, 1981; Friesen *et al.*, 2001, Cooke and Scharenborg, 2008; Scharenborg, 2010).

This analysis is based on the confusion matrices (CMs) for vowel and consonants, which can be used to derive CMs for AFs. For example, a degraded classification of voicing would result in higher confusions between voiced and unvoiced phonemes (e.g., /p/, /b/), while phoneme pairs that differ in the place of articulation (/p/, /d/) would still be distinguishable. A CM for confusions between voiced and unvoiced sounds can be derived by grouping the phonemes from the consonant CM according to the values of the articulatory features shown in Table III (if the presentation was /b/ and the response was /p/, this would increase the voiced-unvoiced confusions in that matrix). Note that this analysis aims at consonant recognition. The phonemes /l/, /ʃ/ and /ts/ were excluded because they would have required the introduction of new feature values for which only few representatives exist.

Instead of reporting the complete CMs for each articulatory feature, we calculated the relative transmitted information T_r for each CM. This measure is comparable to the overall recognition score for that feature, but corrects for chance performance and also takes the distribution of feature values into account (which is important when feature values are not equally distributed for articulatory features). The *absolute* information transmission (or mutual information) is computed using the expression

$$T(x,y) = -\sum_{i,j} p_{ij} \log \frac{p_i p_j}{p_{ij}}, \qquad (1)$$

with the input variable x and the output variable y, each having the possible values $i=1,2,\ldots,k$ and $j=1,2,\ldots,m$, respectively, with the corresponding probabilities p_i, p_j, and the joint probability p_{ij}. The indices i and j refer to the index of the corresponding feature as listed in Table III, or to the consonant index, respectively. The probabilities p_i and p_j are the *a-priori* and *a-posteriori* probabilities for the stimuli, while p_{ij} is a matrix element of the confusion matrix, either of the consonant confusions or the derived matrices for articulatory features. To compare the transmitted information of different features, we report the *relative* information transmission $T_r(x,y)=T(x,y)/H(x)$ with the source entropy $H(x)=\Sigma_i p_i \log(p_i)$ throughout this study (Miller and Nicely, 1955).

2. Spectral distance

Differences in recognition rate may be caused by spectral, temporal or spectro-temporal cues that are associated with the according phoneme. We propose two measures that rely on the spectral properties of speech, in order to investigate to what extent recognition rates and phoneme confusions can be explained by spectral cues. For this approach, we assume that the recognition rate of a phoneme can be modeled by the spectral difference of the masking noise and the phoneme spectrum $D(X_i,N)$, i.e., is it assumed that higher recognition rates are obtained for phonemes with spectral energy above the noise floor. The second assumption is that phonemes are more likely to be confused when their spectra are similar. We propose the spectral inter-phoneme distance $D(X_i,X_j)$ as a measure to quantify this similarity, which can be compared to the corresponding error rate. The level distance between phoneme and masking spectrum $D(X_i,N)$ is defined as

$$D(X_i,N)=\frac{1}{M}\sum_{f,X_i(f)>N(f)+10\ \mathrm{dB}}^{M}(X_i(f)-N(f))^2, \tag{2}$$

where $X_i(f)$ is the long-term frequency spectrum of the ith central phoneme in dB, $N(f)$ is the masking frequency spectrum and M is the number of samples of X_i. To account for the higher critical bandwidth toward higher frequencies in the human auditory system, the long-term spectra are grouped in 45 mel-frequency bins and converted to a dB-scale before calculating the difference between signal and noise. Therefore, level and frequency perception of the human auditory system are approximated, so that the spectral level distance can be seen as a very coarse model for the psycho-physical distance of sounds. The calculation of spectra is described in Section III A. The level of the masking spectrum is raised by 10 dB before the parts of the signal above noise level are used to calculate $D(X_i,N)$. This procedure is similar to the calculation of the articulation index (French and Steinberg, 1947) where the dynamic range of speech sounds (i.e., approx. 30 dB) is adjusted to the mean noise level so that the information-carrying peak energy portions of speech are adjusted to the average noise level.

As a measure for spectral similarity of phonemes, we define the distance between the long-term spectra X_i and X_j of the ith and jth phoneme as

$$D(X_i,X_j)=\frac{1}{M}\sum_{f}^{M}(X_i(f)-X_j(f))^2, \tag{3}$$

where f represents the index of the M frequency bins. By relating those differences to recognition results or error rates, the effect of such dissimilarities can be quantified.

IV. RESULTS

A. Overall recognition scores

Overall recognition accuracies are reported for test Sets *RES*, *DA* and *SNR* in Fig. 2. Scores are broken down into consonant/vowel recognition and the varied parameter. For Sets *RES* and *DA*, the overall recognition rate is about 74%, with large differences between consonants and vowels, the latter producing higher accuracies at this masking level of −6.2 dB.

Recognition scores depending on speech-intrinsic variation obtained with Set *RES* are shown in the left panel of Fig. 2. Best overall results are obtained for high speaking effort (condition ‘loud’, 79.3%) and the reference condition (78.6%). For the categories ‘slow’ and ‘question’, the relative phoneme error rates increased by 7% and 11%, respectively, compared to the reference condition. The absolute performance for ‘fast’ (72.3%) and ‘~~slow~~ (63.3%) is considerably lower than for the reference condition. The relative increase of phoneme errors compared to normal speaking amounts to 29% (fast) and 71% (~~slow~~). The overall scores for categories ‘fast’, ‘soft’, and ‘question’ were significantly different from the reference condition according to McNemar’s test, while scores for slow and loud speaking style were not. However, a further analysis of scores showed that these speaking styles result in significantly different consonant and vowel recognition rates.

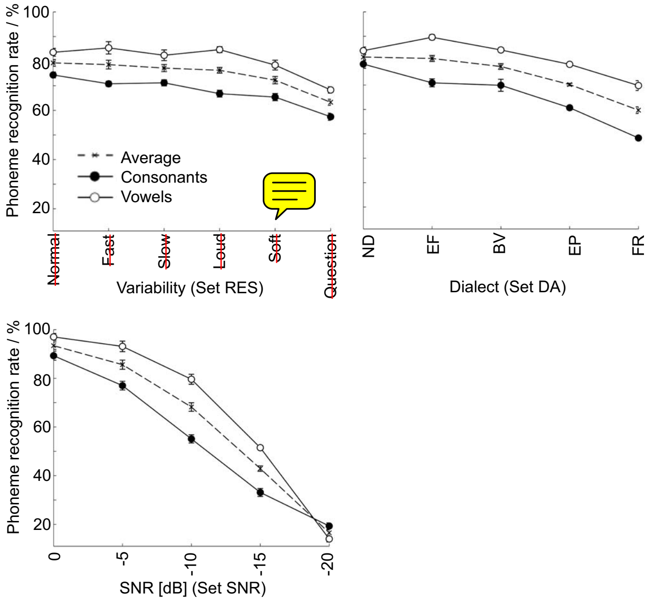

FIG. 2. Phoneme recognition results (% correct) with standard errors, depending on speech-intrinsic variabilities such as speaking rate and style (Set *RES*, left panel) and dialect (Set *DA*, middle panel), and on additive masking noise (right panel). Results for Sets *RES* and *DA* were obtained in listening experiments at −6.2 dB SNR in speech-shaped noise. Variabilities are sorted by average recognition accuracies, which are broken down into consonant and vowel scores.

The overall results (Fig. 2) showed that speech-intrinsic variability induces strong differences in performance for the chosen signal-to-noise ratio: For measurements with varied dialect (Fig. 2, middle panel), the reference condition produces the highest intelligibility (81.5%), as expected for this group of listeners that came from a region without any strong accent. The scores obtained with dialectal or accented speech result in significantly worse scores compared to the reference, with the exception of East Frisian, for which no significant differences are observed when comparing overall results. French accent yields the lowest intelligibility (59.7%), both for consonant and vowel recognition. This corresponds to a relative increase of phoneme error rates of 120%. Even if problematic phonemes that are absent from French are excluded from the analysis, the scores are still below the performance of other conditions. SNR-dependent recognition performance is shown in Fig. 2 (right panel). Vowel accuracies are consistently higher than those of consonants, with the exception of the lowest SNR (−20 dB), which is presumably a result of ceiling effects.

B. Effects of additive noise and intrinsic variability

Since all measurements are based on the same speech corpus, effects of different sources of variability can be expressed in terms of differences of the signal-to-noise ratio that were measured with Set *SNR*. This is shown in Fig. 3 where the accuracies for Sets *RES* and *DA* are projected on the SNR dependent recognition scores. The projection of

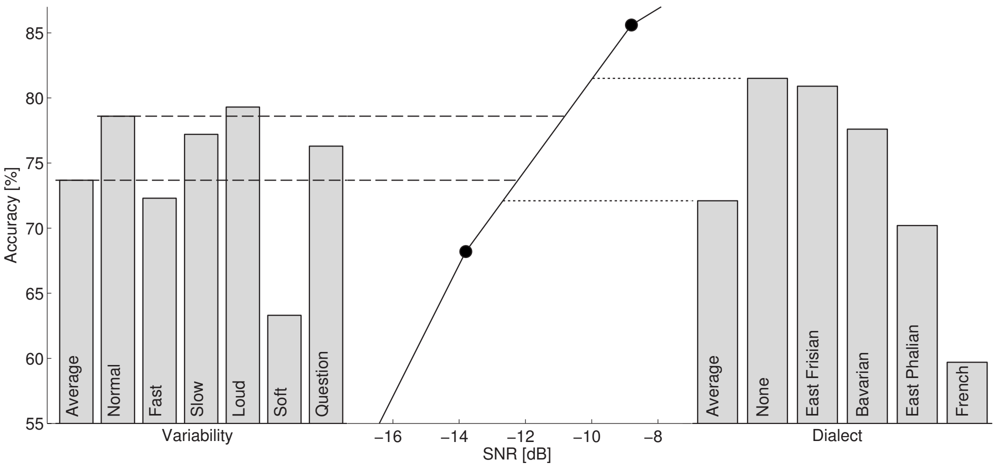

FIG. 3. Average recognition rates depending on speaking variability (left panel), SNR (middle panel) and dialect or accent (right panel). The dashed horizontal lines show the difference between logatomes in the ‘normal’ condition and the average performance of the remaining variabilities. Dotted lines denote differences between the ‘no dialect’ condition and the remaining dialects. By projecting these differences on the middle panel, changes in speaking variability may be expressed in terms of SNR.

variability-dependent scores shows that an average performance corresponds to an SNR of −12.2 dB. The accuracy for the normal speaking style is higher and corresponds to an SNR of −10.8 dB, resulting in a SNR difference of 1.4 dB. In the case of varied dialect or accent, the SNR shift amounts to 2.7 dB. If accuracies obtained with French speech are excluded from this comparison (due to phonetic dissimilarities in German/English vs. French), the gap reduces to 1.5 dB SNR.

Even though these average differences across variability-specific accuracies of 1.4 dB and 2.7 dB (or 1.5 dB), respectively, are comparatively small, the maximum deviation of a specific speaking style from the average is considerable (which is also reflected in the standard deviation of this average): For Set *RES*, the largest performance difference is observed between loud and soft speaking effort, corresponding to an increase of the masking level by 4.2 dB. The standard deviation amounts to 6.2 dB. For Set *DA*, the largest difference occurs between the reference condition and the French accent condition, which corresponds to a 5.5 dB increase of the masking level. The resulting standard deviation amounts to 9.4 dB.

The phoneme confusions obtained with Set *RES* are presented in the consonant and vowel confusion matrices (Table IV and Table V, respectively). The consonant and vowel confusions are reported separately, since for the chosen testing procedure confusions between consonants and vowels do not occur. Results show that the spread in accuracy is larger for consonants (with scores ranging from 36% to 99%) while vowel recognition is more robust in general (72% to 90%). Highest consonant scores were obtained for the phoneme group (/t, s, ʃ, ts, f/), which is in accordance with observations from Phatak and Allen (2007), who found comparable results for the high-scoring consonant phonemes /t, s, z, ʃ, ʒ/. In Table IV, the highest error rates are observed for /a/ and /a:/, which was expected due to their phonetic similarity, as discussed in Section II B, and their spectral similarity, which can be seen from Fig. 1.

TABLE IV. Confusion matrix for consonant phonemes obtained with Set *RES* (−6.2 dB SNR), pooled over all speaking styles, listeners and speakers in this test set. The average recognition rate is 67.7%. Rows (which denote presented phonemes) are normalized and rounded, so that each row adds up to approximately 100% (corresponding to 720 presentations).

	p	t	k	b	d	g	s	f	v	n	m	ʃ	ts	l
p	52.6	2.8	18.2	6.3	1.0	4.3		6.8	6.4	0.6	0.7			0.4
t	0.4	91.4	2.2		4.6			0.1	0.1				1.0	0.1
k	5.8	1.9	67.2	1.4	1.1	16.3	0.1	2.2	1.7	0.8				1.4
b	8.1	0.6	7.1	36.1	4.9	12.2		2.5	19.9	2.2	3.5		0.1	2.9
d	0.7	3.9	1.0	2.9	58.9	12.9	0.1	0.4	2.6	6.8	0.8			8.9
g	1.5	0.7	6.1	6.0	3.8	62.4		1.3	10.1	2.5	1.0			4.7
s		0.3					97.5	0.4				0.8	1.0	
f	3.2		0.4	0.6			12.2	77.2	5.3			0.6	0.4	0.1
v	2.6	1.0	2.2	10.8	3.1	6.0	1.0	7.1	55.3	1.3	4.2		0.3	5.3
n	0.1	0.7	0.4	1.8	7.1	1.7		0.3	3.2	50.6	10.8			23.3
m	1.0	0.1	0.6	6.0	2.8	3.5		1.1	8.6	13.2	48.1			15.1
ʃ	0.1	0.6	0.1			0.1	0.3			0.1		98.5	0.1	
ts	0.1	10.0					3.6					0.1	86.1	
l	0.1	0.7	0.7	0.8	6.9	3.8		0.8	3.3	12.1	4.7	0.4		65.6
Sum	76.3	114.7	106.2	72.7	94.2	123.2	114.8	100.2	116.5	90.2	73.8	100.4	89.0	127.8

TABLE V. CM for vowel phonemes, obtained with Set *RES* (SNR −6.2 dB SNR). The average recognition rate is 80.5%. Rows are normalized, with 100% corresponding to 1152 presentations. For a detailed description, see Table IV.

	a	a:	ɛ	e	ɪ	i	ɔ	o	ʊ	u
a	79.7	19.0	0.1				0.9		0.3	0.1
a:	15.5	83.6					0.2		0.2	0.6
ɛ	0.3		77.6	12.0	8.2	0.4		0.2	0.4	0.9
e			1.6	72.0	15.1	9.7		0.7	0.4	0.5
ɪ			2.4	8.4	86.2	1.3		0.5	0.9	0.3
i				2.3	6.1	90.4			0.3	1.0
ɔ	1.9	1.2					84.4	7.4	4.3	0.9
o			0.1	0.5	0.7	0.2	0.8	71.6	10.7	15.5
ʊ			0.3		1.3	0.1	2.0	11.7	77.9	6.7
u	0.1	0.1	0.1	0.1	1.0	2.4		7.6	6.9	81.6
Sum	97.5	103.9	82.2	95.3	118.6	104.5	88.3	99.7	102.3	108.1

C. Influence of spectral differences

1. Phoneme-noise-distance

The dissimilarities between long-term spectra of high-scoring fricatives and the masking noise (Fig. 1) suggest that spectral properties of phonemes might be a good predictor for recognition rates. This hypothesis was tested by calculating the distance between phoneme and masking spectrum $D(X_i,N)$ according to Eq. (2) and by comparing $D(X_i,N)$ to the recognition rates, as shown in Fig. 4 (left panel). Furthermore, large differences between the long-term spectra of vowels are observed (Fig. 1), while consonant spectra exhibit only small differences over a large frequency scale. It was investigated whether this results in systematic recognition differences between the phoneme types.

A Wilcoxon ranksum test indicated that the distance-to-noise measure $D(X_i,N)$ is not significantly different across consonants and vowels. Hence, the correlation of the phoneme-to-noise distance and recognition rate was calculated based on the data obtained from *all* phonemes. An arcsine-transformation (Studebaker, 1985) was applied to the square root of recognition scores before calculating the correlation in order to linearize the data. The analysis was applied to individual scores (gray-shaded data points in Fig. 4) and showed that the distance-to-noise measure $D(X_i,N)$ is a good predictor for phoneme accuracies ($r=0.74$, $p < 0.0001$).

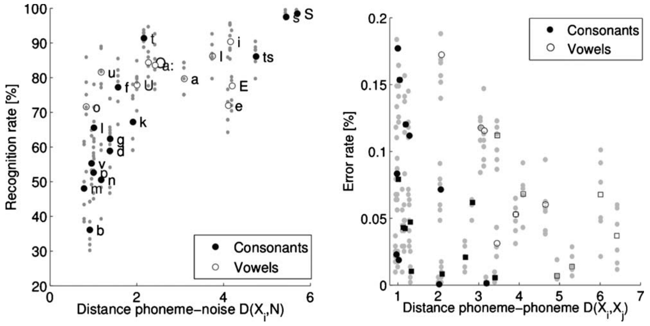

FIG. 4. Left panel: Relation between the phoneme-noise distance and *recognition rates* for consonants and vowels. Next to each data point, the SAMPA transcript of the according phoneme is denoted. The right panel shows the dependency of phoneme-phoneme distance and *error rates* obtained from symmetrized confusion matrices. For each phoneme, several data points are shown which correspond to confusions with 'spectral neighbors', i.e., phonemes that were spectrally closest (marker 'o') and 2nd closest (markers '□') to the presented phoneme. Data points in light gray represent data from individual listeners.

2. Phoneme-phoneme distance

In order to analyze the effect of dissimilarities between *pairs of phoneme spectra*, the spectral inter-phoneme level distance $D(X_i,X_j)$ is compared to *error rates* from confusion matrices. Since $D(X_i,X_j)$ is a symmetric measure (i.e., $D(X_i,X_j)=D(X_j,X_i)$), confusion matrices C were symmetrized by $C_{sym}=\frac{1}{2}(C+C^T)$. The dependency between $D(X_i,X_j)$ and the corresponding error rate is shown in Fig. 4 (right panel). The analysis for *error rates* is limited to confusions to phonemes that were spectrally close to the presented phoneme in order to avoid flooring effects allowing the application of a simple linear model.

In order to test the hypothesis that error rates are related in a different way to phoneme-phoneme distances across consonants and vowels (cf. right panel in Fig. 4), a Wilcoxon rank-sum test was performed, which showed that the phoneme-phoneme distance is significantly higher for vowels than for consonants ($p < 0.0001$). Linear regressions were therefore separately performed for the vowel and consonant group, respectively. The square roots of error rates were subject to an arc-sine transformation before calculating the correlation. As for the phoneme-noise distance, the analysis was based on the data gathered from individual listeners (gray-shaded data points in Fig. 4); it showed that the inter-phoneme distance between spectral neighbors is a good predictor for error rates; the correlation between the two measures was more pronounced for vowels ($r=-0.69$, $p < 0.0001$) than for consonants ($r=-0.61$, $p < 0.0001$).

Even though the very simple measures of spectral phoneme-masker difference and inter-phoneme difference are good predictors for average error rates, they fail to explain the details of the observed recognition and error rates of human listeners. The proposed model relies on spectral super-threshold features, but does not account for purely temporal or spectro-temporal features of the speech signal. An improved prediction requires models that are based on human principles of auditory processing, e.g., the extraction of spectro-temporal features that exhibit a higher signal-to-

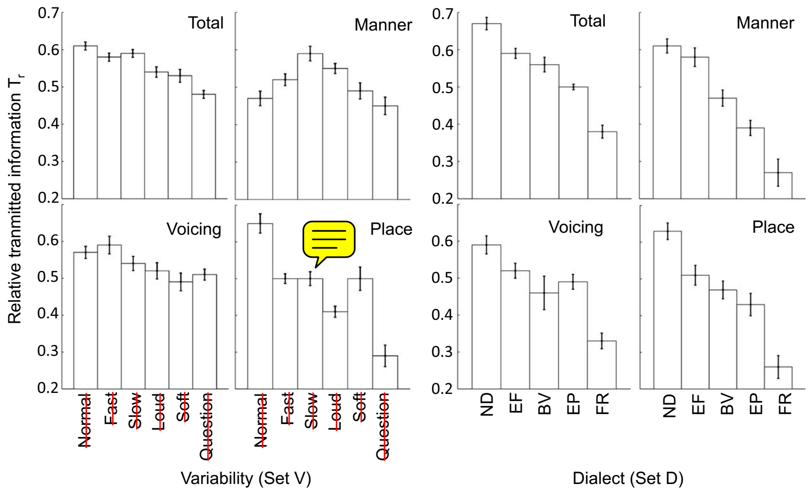

FIG. 5. Relative information transmission T_r depending on speaking variability such as speaking rate and effort (left panel) and dialect and accent (right panel) for selected articulatory features. The error bars denote the standard error across listeners.

noise ratio in an appropriate "glimpse" of the time-frequency distribution (Kleinschmidt and Gelbart, 2002; Barker and Cooke, 2007).

D. Articulatory features and information transmission

The transmitted information of articulatory features is analyzed in order to pinpoint those cues that are most strongly affected in the presence of variabilities. The information channels under consideration were voicing, manner, and place of articulation. These features are well-defined for consonant phonemes, for which the analysis is performed by deriving confusion matrices for articulatory features from the consonant CMs for each variability. These matrices were used to calculate T_r scores. Relative information transmission scores T_r depending on speaking effort, rate and dialect are shown in Fig. 5.

Soft speaking style produced the lowest overall transmission scores, which complies with the consonant results reported in Fig. 2. Major differences compared to the reference condition are high error rates for /p/ and /b/ and confusions between the nasals /n/ and /m/. The latter seems to be the major reason for the low scores of the place feature in soft speaking style (0.29). In contrast to this, place is well recognized for loud speaking style (0.65), with a higher T_r score than normal speaking style (0.50). An analysis of CMs for the place feature showed that this is mainly caused by reduced confusions between anterior and medial placed constrictions of the vocal tract, reflecting overarticulation of loudly spoken utterances.

Slow speaking rate exhibits above average scores for all features. The manner of articulation is particularly well recognized in this case, with a relative increase of 13% of transmitted information, compared to the reference condition. Voicing shows only small variations of T_r scores, which range from 0.49 to 0.59%, and was not found to be significantly influenced by speaking style and rate. This AF therefore appears as being relatively robust toward the discussed variations.

The use of the Oldenburg Logatome Corpus enables an analysis of the same parameters as investigated in related studies (such as the effect of different speaker characteristics on the phoneme scores), but also allows for an analysis of the effect of speech-intrinsic variation based on the same speech material. We analyzed the contribution of these parameters to the observed variance in the recognition results. A repeated measures two-way ANOVA with two across-subjects factors (1. the category for different speaking rates, effort, and style and 2. the choice of speaker) demonstrated significant effects on the relative transmitted information T_r of consonant recognition and articulatory features: both the category for intrinsic variation $[F_1(5,143)=27.8, p<0.001]$ and the choice of speaker $[F_2(4,143)=27.3, p<0.001]$ had a large effect on T_r-scores derived from the consonant confusion matrix.

Similarly, the recognition of manner of articulation was significantly affected by both factors $[F_1(5,143)=11.6, p<0.001, F_2(4,143)=15.1, p<0.001]$, as was the recognition of place of articulation $[F_1(5,143)=47.9, p<0.001, F_2(4,143)=16.9, p<0.001]$. The voicing feature was significantly affected by the choice of speaker $[F_2(4,143)=35.2, p<0.001]$, but not by the category for speaking rates, effort, and style. It therefore seems that the sources of variation analyzed in this study have the largest impact on the place of articulation, while the recognition of the voicing feature seems not to be affected by changes in speaking rate, effort, and style. In all cases, significant interactions between the choice of speaker and the source of variability were observed, indicating that speakers employ individual strategies to produce the desired variation of speech.

A second series of repeated measures ANOVAs was carried out based on the data obtained with Set *DA* with the aim of estimating the influence of such sources of variability on the relative transmitted information. Dialect and accent had a significant effect on the overall T_r-scores $[F_1(4,49)=4.5, p<0.05]$, and on the AFs manner $[F_1(4,49)=4.2, p<0.05]$ and place $(F_1=3.0, p<0.05)$. The recognition of voicing feature was not significantly affected by variability due to dialect and accent.

Relative information transmission depending on dialect and accent is shown in Fig. 5 (right panel). Not surprisingly, the highest values for all features are obtained for standard German. Compared to this, transmitted information for all features is approximately halved for French-accented speech, while the scores for German dialects are in between those conditions. Speech with dialect/accent exhibits the highest relative degradation of information associated with place of articulation. This specific degradation of the place feature is consistent with the notion that the dialects employed differ primarily with respect to the place of articulation. The articulation of voicing and manner, on the other hand, seems to be more constrained by language-specific rules, which results in less variation in transmitted information.

The feature 'manner' is relatively well transmitted for the East Frisian dialect, due to reduced confusions between plosives and fricatives. For example, the error rate for the confusion between /v/ and /b/ is almost halved. In the case of the East Phalian dialect, the voicing feature has relatively high values, as confusions between voiced-unvoiced pairs such as /b/, /p/ are reduced.

V. DISCUSSION

In this study we presented results from speech intelligibility tests with the OLLO logatome speech database that covers several sources of speech-intrinsic variability. From the wide range of sources for such variation in spoken language, we chose those that were found to severely degrade the performance of automatic recognizers. A speech-shaped masking noise was used to avoid ceiling effects in phoneme recognition. While the average degradation in terms of the equivalent loss of the SNR is small (i.e., in the range of a few dB), the degradation is considerable for specific sources of variability: Soft and fast speaking style were identified as the most problematic for human listeners, as relative error rates increased by approximately 70 and 30%, respectively, compared to the reference condition. This reflects the trend identified by Krause (1993) who reported an increase of word error rates for a keyword identification task of approximately 30% when the speaking rate was increased or the speaking effort was lowered. The other sources of variability (reduced rate of speech, increased speaking effort and rising pitch) influenced global recognition scores to a lesser extent (i.e., they resulted in relative changes of error rates between 3% and 11%), but produced shifts regarding the confusion of phonemes.

The presented data analysis is limited to a selection of sources of variability such as speaking rate, effort, style, and dialect and accent. Future experiments may also take other sources of variation into account, such as the effects of age, coarticulation and gender, which has been shown to be a major factor for variations of spoken language (Hazan and Markham, 2004).

A. Comparison with past work

A comparison with important studies on consonant recognition is presented in Fig. 6. It includes data from Phatak and Allen (2007) [PA07], Grant and Walden (1996) [GW96] and Sroka and Braida (2005) [SB05], all of which measured consonant recognition scores in speech-shaped noise. Results from Miller and Nicely (1955) [MN55] who used white noise as masker are also shown. The results obtained in five studies (including the current paper) form three groups with respect to average consonant identification scores: Scores from GW96 and from Sets *RES*, *DA* and *SNR* show good resemblance; the performance obtained in these experiments is between PA07 (for which the performance is 20% higher in average) and SB05 and MN55 (for which it is 20% lower). The highest spread in average recognition performance for sounds masked with speech-shaped noise is observed between PA07 and SB05 with an absolute difference of 39%. Since the slope of the performance-intensity curves for all data given in Fig. 6 is almost identical for 50% consonant intelligibility (~4.5%/dB), the observed difference can be expressed in terms of the SNR: Using a linear interpolation for the mid-region of the performance-intensity curves shown in Fig. 6, the SNR shift was determined which resulted in the smallest rms error between the shifted data from the literature and the scores obtained in this study (Set *SNR*). While this shift was very small for the GW96 data (0.5 dB), the differences for the other studies are more noticeable (PA07: −6.5 dB; SB05: +5 dB; MN55: +6 dB).

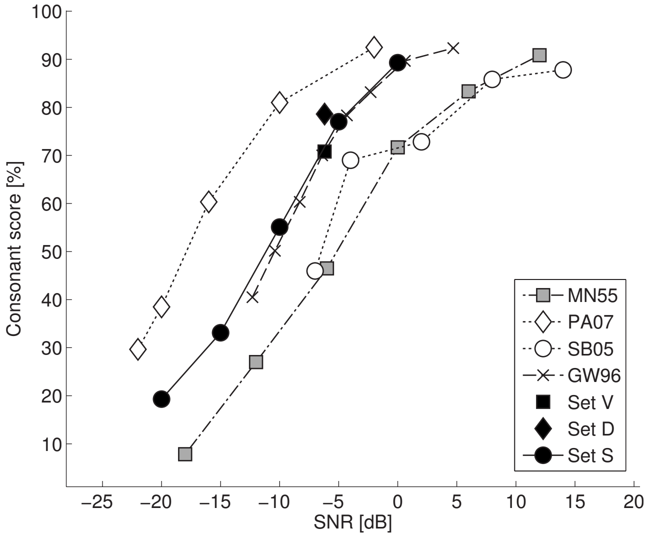

FIG. 6. Comparison of average consonant recognition scores with results from Sroka and Braida (2005) [SB05], Phatak and Allen (2007) [PA07], Grant and Walden (1996) [GW96] and Miller and Nicely (1955) [MN55]. Filled symbols denote results obtained with the OLLO database. Recognition scores for Sets *RES* and *DA* for 'normal' speaking style and 'no dialect' condition include a single SNR and appear as single data points.

There are numerous reasons for the observed variations across studies: Since the spectral difference between phoneme and masker is of primary importance for the phoneme recognition rate (cf. Section IV C and Fig. 4), a major part of the observed variations can be predicted using a simple approach based on the spectral level difference to compensate for the effect of spectral masker and phoneme properties on recognition scores and error rates. The results are in line with findings from PA07 where a modified version of the articulation index (AI) with frequency-dependent weighting coefficients was used, which resulted in a close match of data from MN55, PA07, and GW96. However, the close resemblance of scores between MN55 and SB05 (where masking noises with different characteristics were employed) or the large SNR-shift between PA07 and SB05 (both of which used a speech-shaped masking noise) suggests that average spectral differences are not sufficient as the only important factor in phoneme recognition. Hence, other experimental parameters have to be considered that differ across the studies under consideration: The number of consonant phonemes lies between 12 (SB05) and 18 (GW96). This number influences both chance performance as well as any phoneme confusions that depend on the similarity of phonemes (c.f. Section IV C), which affects overall error rates. Differences across studies may also arise from an unequal number of talkers or listeners, e.g., two talkers and three listeners have been used in (SB05), while in this study, the number of talkers is between one and ten, and the number of listeners between five and ten (depending on the HSR test set, cf. Table II). The difference in SNR calculation might also considerably contribute to the observed shift: For example, the exact definition of the SNR was found to produce differences of more than 3 dB in this study (cf. Appendix C). Finally, differences between PA07 and the other studies may arise from the fact that PA07 removed high-error sounds (which produced error rates > 20% for the quiet condition), which would raise the overall score. Such utterances were not removed for measurements with the OLLO corpus (for which an average recognition score of 0.5% was reported in (Meyer *et al.*, 2006).

Recognition scores depending on speaking rate and style are consistent with other studies. Krause and Braida (2002) presented experiments with conversational and clear speech (i.e., speech with higher intelligibility than conversational speech) with different speaking rates and styles. In our study, we confirm the finding that loudly spoken utterances result in highest intelligibility (after compensating for different absolute speech levels), followed by slow, fast and soft speaking style (in that order). The absolute differences of recognition scores reported by Krause and Braida (2002) are larger than found in this study, i.e., the difference between loud and soft speaking style amounts to 27 percentage-points in Krause and Braida (2002) and to 16 percentage-points in this study. For this comparison we refer to results obtained with conversational speech in Krause and Braida (2002), rather than clear speech that was produced by trained speakers, since speakers recorded for the OLLO database were encouraged to speak in a normal or natural way. However, this larger difference in speech intelligibility score across studies can be explained by the presumably steeper performance-intensity curve for Krause and Braida (2002) where listeners had to identify key words from sentences, in comparison to the flat curve for phonemes employed here (~4.5%/dB, see above). Another factor that significantly influences intelligibility is the inter-individual difference of talkers. Krause and Braida (2004) have shown that two talkers who were trained to produce clear speech at normal speaking rates employed very different strategies for performing this task. For example, large differences of acoustic properties such as voice-onset time and the duration and extent of formant transitions were observed for the talkers. This result underlines the difficulties that arise when results obtained with different speakers are compared, especially when variabilities of speech are considered in connection with unnatural articulation modes (such as speaking "loud" or "clear") where stronger changes due to additional variations (e.g., speaking rate or style) are expected than in normal speech. In this work, we tried to control for these differences by recording several variabilities from the same set of speakers.

B. Comparison between HSR and ASR

In other studies the OLLO corpus has been successfully applied to the problem of ASR (Wesker *et al.*, 2005; Meyer *et al.*, 2007), as an evaluation tool for speech models (Jürgens *et al.*, 2007) and to study speaker discrimination of cochlear implant users (Mühler *et al.*, 2009). By making the speech corpus available for research in HSR and ASR, we hope to promote research dealing with the impact of speech-intrinsic variabilities on both human and automatic recognition. The HSR scores presented in this study may serve as baseline for experiments that aim at narrowing the gap between ASR and HSR, which is still one of the most important challenges in speech research. The speech database, measurement results and detailed results from the analysis can be obtained at http://medi.uni-oldenburg.de/ollo for research purposes.

Bronkhorst *et al.* (1993) showed that the recognition performance increases when meaningful CVCs are presented instead of nonsense CVCs. Such an increase is therefore expected when analyzing continuous conversational (i.e., meaningful) speech instead of logatomes employed in the current study. However, 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. Therefore, 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., Schriberg *et al.* (1984) have shown for transcription of children's speech.

By relating recognition scores obtained for different sources of variability and various SNRs, effects of changed speaking style were expressed in terms of SNR changes. Naturally, these results are valid for medium speech intelligibility only, as for very high SNRs a degradation of 2 dB will have a minor impact on performance, while stronger degradations are obtained when speaking style or dialect is varied (Meyer *et al.*, 2006).

In future research, the impact of intrinsic variation on automatic speech recognition will be assessed and compared to the results obtained with human listening experiments. Such a comparison has been performed earlier (Lippmann, 1997; Sroka and Braida, 2005; ten Bosch and Kirchhoff, 2007; Cooke and Scharenborg, 2008) with the aim of quantifying the gap between HSR and ASR, and the ultimate goal of bridging this gap (i.e., improving ASR) by employing principles that are at work in the human auditory system. While in other studies the focus was laid on extrinsic factors that severely degrade ASR (such as, e.g., the influence of cut-off frequencies of high- and lowpass filtered maskers or the non-stationarity of masking noises) we hope to highlight weaknesses of current ASR systems when speech with intrinsic variation represented in the OLLO speech corpus is to be recognized. The results may then be used to improve the robustness of ASR systems against such variation.

VI. SUMMARY AND CONCLUSIONS

The most important conclusions from this work can be summarized as follows:

(1) The Oldenburg Logatome speech corpus (OLLO) was introduced, and results from human listening tests were reported in terms of error rates and transmission rates of characteristic speech features. The database consists of simple VCV and CVC utterances and covers several speech-intrinsic variabilities. It is available for research purposes for human and automatic speech recognition.

(2) Speech-intrinsic variabilities such as speaking rate, effort and style, and dialect affect the recognition performance of human listeners. High speaking effort produces increased intelligibility and a better transmission of place-of-articulation information compared to normally spoken logatomes, while fast speaking rate or soft speaking style results in increases of phoneme errors by 30% and 70%, respectively (even if the effect of speech level was compensated for). Speech with dialect or accent results in an increase of the error rates by up to a factor of two. The *average* relative increase of error rates due to intrinsic variations was found to be 23% (rate, effort, and style) and 51% (dialect and accent).

(3) To better quantify the effect of intrinsic variabilities, a direct comparison to extrinsic variations (i.e., change in SNR) is possible because the same listeners and speech materials are employed. The presence of varied speaking rate, effort or style corresponds in average to a decrease in SNR of 1.4 dB for a stationary, speech-shaped masking noise (assuming medium speech intelligibility). For dialect and accent, the equivalent decrease in SNR was found to be 2.7 dB. For each individual intrinsic variability, the equivalent SNR degradation amounts to up to 5.5 dB, with standard deviations across variabilities of 6.2 dB (for rate, effort or style) and 9.4 dB (for dialect and accent), respectively.

(4) The analysis of consonant scores based on articulatory features (AFs) showed that the place of articulation is the least robust AF for the variabilities analyzed in this study. On the other hand, the recognition of voiced vs. unvoiced sounds was less affected by changes in speaking style, effort and rate. The strongest effects on consonant recognition and the classification of articulatory features (for the experiment with varying rate, effort and style) resulted from the choice of speaker and the intrinsic variations, while the choice of listener induced only small but significant effects.

(5) The phoneme recognition rate was found to correlate with a simple measure of spectral distance to the masking noise ($r=0.74$), i.e., the spectral characteristics of the masker play an important role in phoneme recognition, which is in line with earlier studies. We also observed that error rates are significantly related to the properties of those alternative phonemes that are spectrally close. This effect was found to be slightly stronger for vowels ($r=-0.69$) than for consonants ($r=-0.61$).

(6) While consonant recognition scores reported here coincide well with data from Grant and Walden (1996), differences of up to 12 dB were found across studies in terms of the SNR corresponding to 50% intelligibility (Miller and Nicely, 1955; Sroka and Braida, 2005; Phatak and Allen, 2007). Our findings of correlations between recognition rates and phoneme-noise distance can account for parts of these differences [and hence confirm findings of Phatak and Allen (2007)]. However, more factors (such as, e.g., the number of response alternatives, the number of phonemes and coarticulation effects in the presented speech items, and the selection and speaking style of the speaker) obviously contribute to the differences across studies. The Oldenburg Logatome Corpus employed here avoids some of these (unwanted) variability effects by using a fixed word format and providing a number of different speaking styles with the same respective talker. It therefore produces phoneme scores that are in between the extreme high and low scores found in the literature.

ACKNOWLEDGMENTS

Supported by the DFG (Grant No. SFB/TRR 31, 'The active auditory system'; URL: http://www.uni-oldenburg.de/sfbtr31). The OLLO speech database OLLO has been developed as part of the EU DIVINES Project No. IST-2002-002034. We thank Matthias Wächter and Markus Röhl for conducting several of the listening experiments. We also thank Alfred Mertins, Jörn Anemüller, Kirsten Wagener, Stéphane Dupont, and Christophe Ris for their support and contributions to this work.

APPENDIX A: ASR EXPERIMENTS FOR PHONEME SELECTION

An ASR experiment was performed with the aim of identifying the consonants that result in high error rates and therefore should be included in the OLLO database. Spectro-temporal ASR features (Kleinschmidt and Gelbart, 2002) served as input to a non-linear neural network (multi-layer perceptron, MLP) that was trained and tested using a phoneme-labeled speech database (TIMIT). Results were analyzed on a frame-by-frame basis and phonemes were sorted by their relative error rate. The corresponding confusion matrix is shown in Fig. 7. The phonemes that were selected based on this experiment are highlighted in the figure.

APPENDIX B: ASR EXPERIMENTS FOR SPEAKER SELECTION

Since HSR experiments could not be performed with the complete OLLO database, test sets were compiled to investigate different sources of variability. Earlier studies have shown that the intelligibility of speech strongly depends on the choice of speaker (Barker and Cooke, 2007). Hence, a

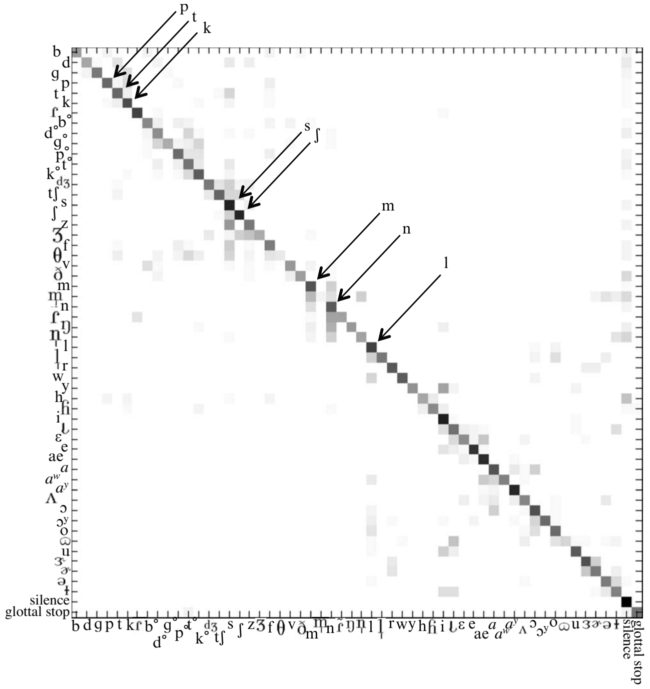

FIG. 7. Phoneme confusion matrix obtained in an ASR experiment. In this row-normalized CM, black color denotes unity and white color corresponds to chance performance. The eight consonant phonemes that were selected to be included in the OLLO database based on this experiment are marked with arrows.

selection procedure was carried out to avoid the use of speaker data that produces very high or low scores compared to those for 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 Mel-frequency cepstral coefficient (MFCC) features and a Hidden Markov Model classifier. This procedure was performed for all speakers in the corpus; the four speakers selected for Set *RES* produced an average score in the same range as the average score measures for all speakers without dialect (84.3% vs. 84.1%). Similarly, the scores of ten speakers (one male and one female speaker from each dialect region) included in Set *DA* were comparable to the score averaged over all 50 speakers (75.4% vs. 75.8%, respectively).

APPENDIX C: SNR CALCULATION

For such short utterances as logatomes, the adjustment and interpretation of the SNR is not a trivial issue because the short-term level derived from each logatome varies considerably across logatomes even if exactly the same recording conditions are used (i.e., technical conditions, speaker, speech rate, speech effort, etc.). Obviously, the reliability of the short-term level as an estimate of the "true" speech level decreases with decreasing duration of the speech segment. One option for a more valid speech level measure as an input to the SNR measure would therefore be to use the *average* power of all speech samples in the database, since the long-term SNR has been shown by the Articulation Index and the Speech Intelligibility Index to be a reliable measure for *average* speech intelligibility. Using such a long-term speech level, changes in the recording conditions (e.g., variations of the distance between speaker and microphone) can be reliably detected and compensated for. On the other hand, the short-term rms level of a *single* utterance is an easily computable local measure that does not rely on the (statistical) properties of the remaining speech corpus and captures best the properties of the individual speech item. Hence, the short-term SNR is very popular in speech research and has been used, e.g., in other studies that make use of CV utterances in noise (as, e.g., Cooke and Scharenborg, 2008). However, due to the large statistical uncertainty with short speech segments, the intelligibility obtained from short VCV and CVC combinations varies considerably across speech items in a way not predictable from the variability of the short-term SNR and only partially predictable from the long-term SNR (Kollmeier, 1990), which compensates for slow variations of the recording conditions. Since these variations were already controlled and compensated for during the recording of the OLLO speech corpus and for the sake of simplicity and compatibility with recent studies, we used the short-term SNR derived from each single utterance throughout this study.

For measurements with Sets *DA* and *SNR*, the SNR was calculated by relating the root-mean-square (rms) value of the speech segments of each audio signal and the rms value of a masking noise of equal length. A simple voice detection algorithm based on an energy criterion was used to extract connected speech segments. Random control samples were chosen to control proper functioning of that algorithm. For utterances from Set *RES*, a different SNR calculation scheme was applied: In this case, the rms levels of the whole utterance (including silence) and a noise segment of equal length were used to adjust the SNR. Since the length of silence before and after each logatome is 500 ms and because the variation of temporal spread of identical logatomes is relatively small, this corresponds to 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.

Allen, J. B. (**1994**). "How do human process and recognize speech?," IEEE Trans. Speech Audio Process. **2**, 567–577.

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

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

Chang, S., Wester, M., and Greenberg, S. (**2005**). "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," Speech Commun. **47**, 290–311.

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

Cooke, M. P., Green, P. D., Josifovski, L. B., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and uncertain acoustic data," Speech Commun. **34**, 267–285.

Dreschler, W. A., Ludvigson, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speechlike spectral and temporal properties for hearing instrument assessment," Audiology **40**, 148–157.

Dubno, J. R., and Levitt, H. (**1981**). "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**, 249–261.

Fissore, L., Mertins, A., Ris, A., Rose, R., Tyagi, V., and Wellekens, C., (**2007**). "Automatic speech recognition and speech variability: A review," Speech Commun. **49**, 763–786.

Flege, J. E., Schirru, C., and MacKay, I. R. A. (**2003**). "Interaction between the native and second language phonetic subsystems," Speech Commun. **40**, 467–491.

Fosler-Lussier, E., and Morgan, N. (**1999**). "Effects of speaking rate and word frequency on conversational pronunciations," Speech Commun. **29**, 137–158.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Gelfand, S., Piper, N., and Silman, S. (**1985**). "Consonant recognition in quiet as a function of aging among normal hearing subjects," J. Acoust. Soc. Am. **78**, 1198–1206.

Grant, K. W., and Walden, B. E. (**1996**). "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am. **100**, 2415–2424.

Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.

Hermansky, H., and Morgan, H. (**1994**). "RASTA processing of speech," IEEE Trans. Speech Audio Process. **2**, 578–589.

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

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

Kleinschmidt, M., and Gelbart, D. (**2002**). "Improving word accuracy with Gabor feature extraction," in Proceedings of the International Conference on Spoken Language Processing (ICSLP), pp. 545–548.

Kliem, K. (**1993**). "Entwicklung und Evaluation eines Zweisilber-Reimtestverfahrens in deutscher Sprache zur Bestimmung der Sprachverständlichkeit in der klinischen Audiologie und Nachrichtentechnik (Development and evaluation of a German bisyllabic rhyme test for speech intelligibility measurements in clinical audiology and communications engineering)," Ph.D. thesis, University of Oldenburg, Oldenburg, Germany.

Kohler, K. (**1995**). *Einführung in die Phonetik des Deutschen (Introduction to German Phonetics)* (Erich Schmidt, Berlin).

Kollmeier, B. (**1990**). "Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache (Measurement, modeling and improvement of speech intelligibility)," Habilitation thesis, University of Göttingen, Fachbereich Physik, Göttingen.

Kollmeier, B., Kliem, K., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am. **102**, 2412–2421.

Kollmeier, B., and Wallenberg, E.-L. (**1989**). "Sprachverständlichkeitsmessungen für die Audiologie mit einem Reimtest in deutscher Sprache: Erstellung und Evaluation von Testlisten (Speech intelligibility measurements for audiology based on a German rhyme test: Preparation and evaluation of test lists)," Audiologische Akustik **28**, 50–65.

Krause, J. C. (**1993**). "The effects of speaking rate and speaking mode on intelligibility," Master's thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Krause, J. C., and Braida, L. D. (**2002**). "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**, 2165–2172.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378.

Li, C.-n. (**2003**). "Accent, intelligibility, and comprehensibility in the perception of foreign-accented Lombard speech," J. Acoust. Soc. Am. **114**, 2364.

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

MacArthur, T. (**1992**). *The Oxford Companion to the English Language*, Oxford University Press, New York.

Meyer, B. T., Wächter, M., Brand, T., and Kollmeier, B. (**2007**). "Phoneme confusions in human and automatic speech recognition," in Proceedings of Interspeech, pp. 1485–1488.

Meyer, B. T., Wesker, T., Brand, T., Mertins, A., and Kollmeier, B. (**2006**). "A human-machine comparison in speech recognition based on a logatome corpus," in Proceedings of the Workshop on Speech Recognition and Intrinsic Variation, pp. 95–101.

Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some english consonants," J. Acoust. Soc. Am. **27**, 338–352.

Mühler, R., Ziese, M., and Rostalski, D. (**2009**). "Development of a speaker discrimination test for cochlear implant users based on the OLLO logatome corpus," ORL **71**, 14–20.

Müller, C. (**1992**). "Perzeptive Analyse und Weiterentwicklung eines Reimtestverfahrens für die Sprachaudiometrie (Perceptual analysis and development of a ryhme test for speech audiometry)," Ph.D. thesis, Georg-August-Universität, Göttingen, Germany

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

Scharenborg, O. (**2010**). "Modeling the use of durational information in human spoken-word recognition," J. Acoust. Soc. Am. **127**, 3758–3770.

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

Siegler, M. A., and Stern, R. M. (**1995**). "On the effect of speech rate in large vocabulary speech recognition systems," in Proceedings of ICASSP, pp. 612–615.

Siniscalchi, S. M., Svendsen, T., and Lee, C.-H. (**2008**). "Towards a detector-based universal phone recognizer," in Proceedings of ICASSP, pp. 4261–4264.

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," *Automatic Speech and Speaker Recognition*, edited by C.-H. Lee, F. K. Soong, and K. K. Paliwal (Springer, Berlin), Chap. 15.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Tchorz, J., and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition," J. Acoust. Soc. Am. **106**, 2040–2050.

ten Bosch, L., and Kirchhoff, K. (**2007**). "Bridging the gap between human and automatic speech recognition," Speech Commun. **49**, 331–335.

Wang, M., and Bilger, R. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.

Weintraub, M., Taussig, K., Hunicke-Smith, K., and Snodgrass, A. (**1996**). "Effect of speaking style on LVCSR performance," in Proceedings of the Addendum of ICSLP, pp. 1457–1460.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (**2005**). "Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines," in Proceedings of Interspeech, 1273–1276.