Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines

Thorsten Wesker[1], *Bernd Meyer*[1], *Kirsten Wagener*[1], *Jörn Anemüller*[1]
Alfred Mertins[2], *Birger Kollmeier*[1]

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

Abstract

This paper introduces the new OLdenburg LOgatome speech corpus (OLLO) and outlines design considerations during its creation. OLLO is distinct from previous ASR corpora as it specifically targets (1) the fair comparison between human and machine speech recognition performance, and (2) the realistic representation of intrinsic variabilities in speech that are significant for automatic speech recognition (ASR) systems. To enable an unbiased human-machine comparison, OLLO is designed for recognition of individual phonemes that are embedded in logatomes, specifically, three-phoneme sequences with no semantic information. A balanced set of target-phonemes important for human and automatic speech recognition has been chosen, drawing on pilot ASR studies and cross-fertilization from the field of human speech intelligibility testing. Several intrinsic variabilities in speech are represented in OLLO, by recording from 40 speakers from four German dialect regions, and by covering six articulation characteristics. Results from preliminary phonetic time-labeling and ASR experiments are promising and consistent with corpus variabilities.

1. Introduction

The OLdenburg LOgatomes (OLLO) speech corpus has been developed as part of the EU project DIVINES (Diagnostic and Intrinsic Variabilities in Natural Speech). One aim of the project is a better understanding of human speech recognition and thereby gaining benefit in developing machine recognizers. DIVINES considers natural speech variabilities that are not influenced by the surrounding, but could lower the performance of Automatic Speech Recognition (ASR), e.g., variability due to speaker dialect or articulation. The speech database presented in this paper is specially designed for better modeling, feature extraction and adaptation in the presence of intrinsic variabilities. It is suitable for both ASR experiments and speech intelligibility tests with humans.

While ASR has seen many advances in recent years, the error rates of machines are still an order of magnitude higher than those of humans. This suggests that human recognition uses a wide spread set of cues for comprehension and error correction, which are not yet accessible for machines. In order to conduct a fair human-machine comparison both sides should utilize the same kind of information. All information that could give an advantage to only one side must be eliminated. Former studies show that error rates converge in testing setups where humans could not use supplementary cues, such as the context of the conversation, the grammar of the spoken language and certain words [1]. Hence one should perform the recognition tests on the smallest information carriers in natural speech, so that the additional cues, which are usually exploited by humans, are reduced to a minimum.

If the utterances carry no meaning and the testing is done on the smallest distinct speech entities, only the clean neurosensory process of hearing and recognition, without further cognitive processing and error correction, is considered and rated. By this it should be possible to identify those physical features, which are at least needed to recognize and to distinguish each utterance. Testing on phonemes would be appropriate for this aim. Therefore logatomes (greek: "cutout of word") consisting of three phonemes were chosen as testing material for the OLLO speech corpus.

2. Choice of Phonemes

Each logatome in the OLLO speech database consists of three phonemes. The relevant phoneme on which the tests will be performed is embedded between two nearly identical frame phonemes. Thereby it becomes possible to analyze the influence of coarticulation and, in closed intelligibility-tests, humans and machines can choose out of the same range of answers. The structures of the logatomes are either vowel-consonant-vowel (VCV) or consonant-vowel-consonant (CVC), depending on the type of the target-phoneme in the middle.

In order to investigate the influence of variabilities, each logatome was recorded multiple times. To keep the recording time per speaker at a reasonable level, the number of phonemes had to be limited. Phonemes that are critical in either human or automatic recognition of speech were selected, so that significant differences in recognition rates may be obtained with smaller test sets.

2.1. Critical phonemes in human speech recognition

To find those German phonemes which are most critical in recognition or often mixed up by humans, monosyllable and disyllable rhyme tests with normal-hearing listeners were analyzed [2],[3],[7],[6]. The results suggest that the following phonemes should be taken into account:

- consonants: [b] [d] [f] [g] [k] [l] [p] [ɾ] [s] [v]
- vowels: [æ] [ɛ] [ɪ] [i] [ʊ][u] [y]

2.2. Critical phonemes in ASR

In order to determine the phonemes that are most critical in ASR, the matrix of confusion was calculated in a phoneme recognition experiment. Gabor features [4, 5] were used as

input to a non-linear neural network (multi-layer perceptron, MLP). The MLP was trained and tested with features calculated from a phoneme-labeled speech database. The experimental parameters were chosen as follows:

- 60 dimensional Gabor feature vectors with additional delta- and double delta features were used as input to an MLP [1].
- The TIMIT speech database was used as training- and test material (disjoint sets). Noise signals as provided by the Aurora 2 paradigm were applied to the speech data.
- The MLP had 180, 1000 and 56 neurons in input, hidden and output layer, respectively.
- Part of the 61 phonemes in the original TIMIT database were combined according to the ICSI56 phoneme set, so that the labels contained 56 different phonemes.
- The TIMIT database contains English utterances, whereas the OLLO corpus should contain German logatomes. Hence, critical phonemes that are present in German as well as in English language were selected for the corpus.

Phonemes were sorted by their relative error rate. The following phonemes were selected for the corpus because they appear in both German and English language, produced high error rates in the experiment and are often present in phoneme confusions:

- [d] [v] [f] [g] [z] [m] [n] [ʃ]

The first four of them were already included in the set of phonemes that are critical for humans.

2.3. Final Set of Phonemes

In order to keep the number of alternatives equal for both, humans and machines, only the recognition of the middle-phoneme is investigated. Different phonemes were chosen for VCVs and CVCs, so that the number of middle-phonemes can be increased, while the resulting number of logatomes is still small enough to be recorded in a single session. The final phoneme set for VCVs is:

- [a] [ɛ] [ɪ] [ɔ] [ʊ]
- [d] [t] [g] [k] [f] [s] [b] [p] [w] [z] [m] [n] [ʃ] [l]

For CVCs it is:

- [a] [ɛ] [ɪ] [ɔ] [ʊ] [aː] [eː] [iː] [oː] [uː]
- [d] [t] [g] [k] [f] [s] [b] [p]

Combining these phonemes gives a total number of 150 different logatomes.

The German orthographic transcriptions, which could be visually presented to the speakers during recording, were prepared by a phonetician. The following are examples for VCVs:

taht, tuht, teht tatt, tutt, tett

sahs, suhs, sehs sass, suss, sess

pahp, puhp, pehp papp, pupp, pepp

Examples for CVCs:

ollo, oggo, otto elle, egge, ette ullu, uggu, uttu

[1] SPRACHcore / QuickNet software package provided by ICSI, http://www.icsi.berkeley.edu

Figure 1: *Regional provenance of dialect speakers in the OLLO corpus.*

3. Speakers and Variabilities

To enable the development of techniques that will have better capacity in handling intrinsic speech variabilities, the corpus is specially grouped into speaker-independent and speaker-dependent variabilities. In the OLLO corpus there is a total number of **40 speakers**. The **speaker-dependent variabilities** are **gender** (19 male, 21 female), **age** (ranging from 18 to 65) and regional German **dialect**. The dialects are: standard German recorded in Oldenburg (Speakers chosen from the university population), Bavarian (BV) recorded in Munich (speakers chosen from the rural surrounding of Munich), East Frisian (EF) recorded in Oldenburg (speakers chosen from the rural surrounding of Oldenburg), Eastphalian (EP) recorded in Magdeburg (speakers chosen from the suburbs of Magdeburg). 10 speakers from each dialect were recorded. In Fig. 1 their geographical provenance is shown. Each of OLLO's 150 logatomes has been recorded 3 times in 6 different articulation characteristics.

The six **speaker independent variabilities** are: **speaking rate** (fast, normal, slow), **speaking effort** (low, normal, high), **speaking style** (statement, question). Only one of these characteristics is varied at a time, the others remain normal. This results in 2700 recorded logatome items per speaker. In addition, 72 monosyllabic words (part of the monosyllabic rhyme test of Von Wallenberg and Kollmeier [11]) and 20 sentences, (part of the Göttingen sentence test of Kollmeier and Wesselkamp [12]) were recorded with each speaker. Both sets of this subcorpus represent the mean German phoneme distribution and are planned for ASR speaker adaptation or other training purposes. It is known that the type of the microphone influences the performance of ASR systems [1]. So OLLO was recorded simultaneously with four different **microphone** setups (2 x HQ, PC, cell-phone), as another variability, all typical for certain ASR tasks.

4. Recording Setup

4.1. Technical Equipment

All recordings took place in sound-insulated free-field and in sound-insulated audiometry rooms. In the near-field there were installed: a high-quality condenser-microphone (AKG C1000 S), a typical electret desktop PC microphone (Speedlink Pan SL-8704-SSV) and a cell-phone headset-microphone (Nokia

standard). A second high-quality condenser-microphone (AKG C1000 S) was installed in the far field (1.50m from the speaker's position). Studio quality harddisk-recording equipment sampled the raw data. The setup consisted of a RME QuadMic microphone pre-amplifier and an RME Hammerfall AD-Converter with a sampling rate of 44.1 kHz and a resolution of 32 bit. The amplifiers were always adjusted to the same gain settings. A quality check protocol was developed to prevent clipping and inadequate signal to noise ratios. The recording was managed by a specially developed software tool based on MatLab Version 7 (The MathWorks) and SoundMex (HörTech GmbH). The software tool presents the desired logatomes to the speaker and controls the blockwise random recording order as well as the storage of the recorded and digitized speech.

4.2. Recording Conditions

The items to be recorded were presented to the speakers on a computer screen in the orthographic transcription shown at the end of Sec. 2.3. The logatomes were presented in random order, to avoid systematic errors. A complete randomized combination of logatomes and variabilities confused the speakers in preliminary recordings. Therefore, a random order of blocks with fixed variability was chosen (different for each speaker). In each block, the order of logatomes was chosen randomly. The actual variability and logatome was visually presented to the speakers. Between the blocks with different variabilities, a special warning was given to remind the speakers that the variability was changing.

The speakers were introduced into the recordings and advised to speak in a natural manner (i.e., not to suppress their dialect). In a preliminary test, the realization of variabilities was checked and corrected if necessary. In cases where pronunciation was not well defined by the written representation, speakers were recommended a certain pronunciation. The speakers were advised to take as much time and recreations as they needed to gain a constant speech quality. The average duration of the whole recording procedure was 3.5 hours per speaker.

5. Postprocessing of Recorded Material

Since the recording of an item was stopped and the recording of the next item initialized by either pressing a computer mouse button or the space button of a computer keyboard, a click was introduced at the end of the recordings. In order to eliminate clicks that were too close to the utterance itself all files were automatically post-processed with the help of a click detector. In cases where a click was found near or even in the speech signal or an unclear situation occurred, the files were sorted out to be checked manually. To this end, a semi-automatic click-removal tool, developed in MatLab, was employed. Recordings that were incomplete or could not be corrected manually were sorted out. The silence at the beginning and at the end of each recording was limited to 500ms. The click free and cut raw data WAV-files were normalized to 99% amplitude and stored with 16 bit resolution. They were low-pass filtered with 8 kHz cutoff frequency and sampled down to 16 kHz. All necessary meta-information was stored in list files.

5.1. 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), cf. [10]. In a nutshell, the MAUS labeling procedure is similar to HMM forced alignment approaches. 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 [9] and [10].

The high-quality recording with AKG microphone of the OLLO corpus was chosen for labeling. Data was preprocessed as described above. All 150 logatomes were transcribed in the SAM phonetic alphabet (SAM-PA) and the transcription 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. Compensating for result differences that could be accounted for by negligible shifts in phoneme boundary positions, an estimated 1.2% of utterances 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 subjects. 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.

6. Corpus Retrieval

The entire OLLO-Corpus, including a detailed description, wordlists, labeling files, technical specifications and calibration data (normalization coefficients and dB(SPL) values) is open to the public. The corpus is approx. 4.6 GB in size and contains a total of 104,628 files corresponding to 43.3 hours of speech. It can be downloaded for free from **http://sirius.physik.uni-oldenburg.de** or ordered on one DVD via e-mail for a nominal fee.

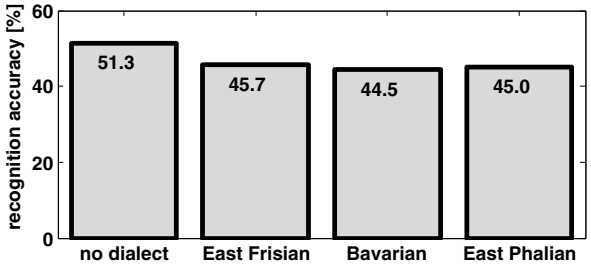

Figure 2: *Recognition rates depending on regional dialect. The ASR frontend was trained on German language without any dialect which contributes to the better performance in the first category.*

7. Preliminary ASR Experiments

Preliminary ASR Experiments were conducted with OLLO using the Loquendo ASR engine which is based on a hybrid approach that combines a Hidden Markov Model and a neural

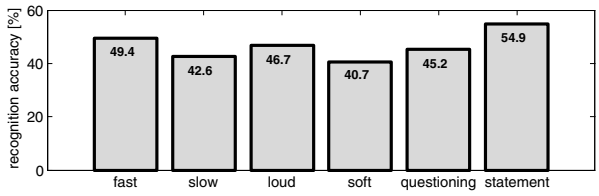

Figure 3: *Recognition rates depending on intrinsic variabilities in speech. The results demonstrate that a change in speaking rate and effort can increase word error rates by up to 31.5 %.*

net. The emission probabilities of the HMM states were estimated by a Multi Layer Perceptron. Decoding of these states is carried out at phonetic level [8]. A phonetic transcription of logatomes was used to investigate the influence of variabilities. Since the ASR system was already trained on German language, all recorded utterances were used as test material.

The total recognition accuracy for the OLLO corpus was 46.59 %. The performance per speaker ranges from 36.6 % to 60.2 % (standard deviation of 5.4 %). The differences in dependency of the speaker's gender are much smaller: Accuracies are 47.2 % and 46.0 % for male and female speakers, respectively. Figs. 2 and 3 show the recognition rates for different dialects and in dependency of variabilities.

In this setup, the regional dialects are harder to recognize than standard German and the normal spoken statements show best results. The confusion matrix for logatomes (Fig. 4) shows that most errors occurred due to substitution of a certain vowel by its own longer or shorter version and difficulties in distinguishing [o] from [u]. Future ASR experimental results will be compared with results received from equivalent human speech intelligibility tests.

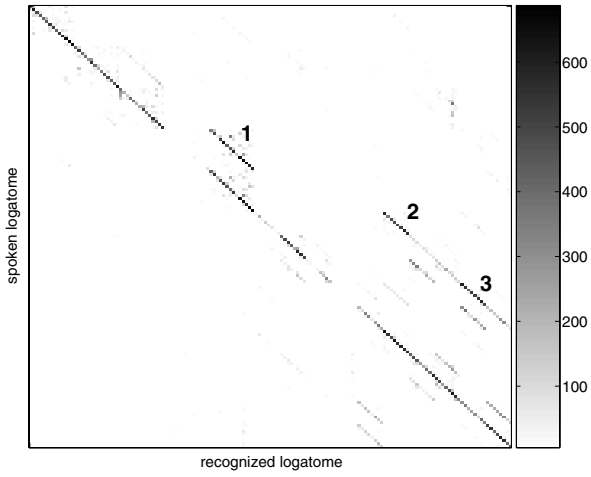

Figure 4: *Matrix of confusion for 150 logatomes. The deviations from the diagonal in area 1 arise from confusions of phonemes [o] and [u] (e.g. "ollo" is recognized as "ullu"). In areas 2 and 3, long and short phonemes are confused (e.g. "sass" is recognized as "sahs").*

8. Summary

The creation of the OLLO speech corpus was described. The purpose of this new database is to perform a wide range of human-machine comparisons in speech recognition under several intrinsic variabilities of speech. The results of the comparisons should lead to a better understanding of speech processing by the human auditory system and thereby enhancing the possibilities and the performance of artificial speech recognizers. First ASR experiments with the database seem promising.

9. Acknowledgements

OLLO has been developed as part of the EU DIVINES Project IST-2002-002034.

10. References

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

[2] Dubno, J.R.; and Levitt H. - "Predicting consonant confusions from acoustic analysis", J. Acoust. Soc. Am. 69 (1): 249-261,1981.

[3] Gelfand, S.A.; Piper, N. and Silman, S. - "Consonant recognition in quiet as a function of aging among normal hearing subjects", J. Acoust. Soc. Am. 78 (4): 1198-1206, 1985.

[4] Kleinschmidt, M. and Gelbart, D. - "Improving word accuracy with Gabor feature extraction", ICSLP, Denver, 2002.

[5] Meyer, B. and Kleinschmidt, M. - "Robust Speech Recognition Based on Localized Spectro-Temporal Features", ESSV, Karlsruhe, 2003.

[6] Kliem, K. - "Entwicklung und Evaluation eines Zweisilber-Reimtestverfahrens in deutscher Sprache zur Bestimmung der Sprachverständlichkeit in der klinischen Audiologie und Nachrichtentechnik", PhD thesis, Universität Oldenburg, 1993.

[7] Müller, C. - "Perzeptive Analyse und Weiterentwicklung eines Reimtestverfahrens für die Sprachaudiometrie", PhD thesis, Georg-August-Universität Göttingen,1992.

[8] Colibro, D.; Fissore, L.; Popovici, C.; Vair, C. and Laface P. - "Learning Pronounciation and Formulation Variants in Continuous Speech Applications", ICASSP, 2005.

[9] Kipp, A. and Wesenick, M.-B. - "Das Münchener AUtomatische Segmentationssystem (MAUS)", Memo-95-95, Institut für Phonetik und Sprachliche Kommunikation, University of Munich, 1995.

[10] Kipp, A.; Wesenick, M.-B. and Schiel, F. - "Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora", Proceedings of the ICSLP 1996, pp. 106–109, 1996.

[11] Kollmeier, B. and Wallenberg, E.-L. - "Sprachverständlichkeitsmessungen für die Audiologie mit einem Reimtest in deutscher Sprache: Erstellung und Evaluation von Testlisten", Audiologische Akustik, 28(2), p. 50-65.

[12] Kollmeier, B.; Kliem, K. and Wesselkamp, M. - "Development and Evaluation of a German Sentence Test for objective and subjective Speech Intelligibility Assessment", Journal of the Acoustical Society of America, 102(4), p. 2412-2421.