Modelling the Human-machine Gap in Speech Reception: Microscopic Speech Intelligibility Prediction for Normal-hearing Subjects with an Auditory Model

Tim Jürgens, Thomas Brand, Birger Kollmeier

Institute of Physics, Carl-von-Ossietzky University Oldenburg, Germany

tim.juergens@uni-oldenburg.de, thomas.brand@uni-oldenburg.de,
birger.kollmeier@uni-oldenburg.de

Abstract

In this study speech intelligibility in noise for normal-hearing subjects is predicted by a model that consists of an auditory preprocessing and a speech recognizer. Using a highly systematic speech corpus of phoneme combinations (logatomes) allows the analysis of response rates and confusions of single phonemes. The predicted data is validated by listening tests using the same nonsense speech material. If testing utterances that are not identical to those in training material are used, the psychometric function in noise is predicted with an offset of 13 dB to higher signal-to-noise-ratios (SNR). This is consistent with the man-machine performance gap between human speech recognition (HSR) and automatic speech recognition (ASR) [1].

However, this offset reduces to 4 dB in a second model design with identical recordings for training and testing. Furthermore predicted confusion matrices are compared to those of normal-hearing subjects with the second model design.

Index Terms: speech intelligibility prediction, auditory model, confusion matrix, phonemes

1. Introduction

Typical models that predict speech intelligibility in noise for normal-hearing subjects, as e.g. the Speech Intelligibility Index (SII) [2], analyse the long-term spectra of speech and noise separately in different frequency channels. The outcome of these models can be transformed to the speech reception threshold (SRT), which gives the SNR of 50% speech intelligibility and the slope of the psychometric function. Recognition rates and confusions of phonemes can not be studied using these models.

The model proposed here is based on an idea of Holube and Kollmeier [3] and consists of a psychoacoustically motivated preprocessing of the time-signal and a standard dynamic-time-warp (DTW) speech recognizer [4]. By determining the distances between a test utterance and training utterances "on a perceptual scale" the utterance with the least distance is taken as the recognized one.

For prediction and validation we used the context-free speech database Oldenburg Logatome Corpus (OLLO) [5]. It contains 70 different vowel-consonant-vowel (VCV) and 80 CVC logatomes composed of German phonemes. Each logatome was recorded 18 times by each speaker. 6 different speech articulation styles are included: "slow", "normal", "fast", "loud", "quiet" and "questioning". The use of this corpus allows systematical investigations of phoneme recognition rates and confusions. At the same time it avoids that human listeners can use any semantic knowledge for intelligibility.

2. Measurements

2.1. Method

10 clinically normal-hearing subjects (7 male, 3 female) aged between 19 and 37 years were employed. The intelligibility of 150 logatomes was measured in a sound isolated booth at different signal-to-noise-ratios. All recordings were taken from the OLLO database and were spoken by a single German speaker with speech variability "normal". The 150 recordings were randomly split into two lists of the same length for each of the 5 SNRs and the resulting 10 lists were randomly inter-leaved for presentation. The speech was presented at a level of 60 dB SPL via Sennheiser HDA 200 headphones that were free-field equalized using a FIR-filter with 801 coefficients. A non-modulated running noise with speech-like frequency spectrum was used (ICRA-1 noise, [6]). All audio signals were presented diotically. Response alternatives for a single logatome had the same preceding and subsequent phoneme (closed test); hence, the subject had to choose from 10 or 14 alternatives when a CVC or a VCV was presented, which one was recognized.
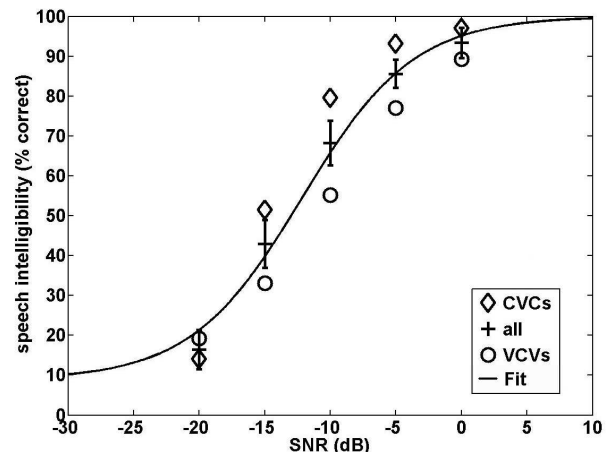
2.2. Results

Figure 1: *Psychometric function for normal-hearing subjects measured with logatomes in ICRA-1 noise at 5 fixed SNR respectively. Error bars show the inter-individual standard deviation for 10 subjects. The fitted function is shown for comparison.*

Figure 1 shows the results of the speech intelligibility test plotted versus the SNR. Every symbol represents the mean intelligibility of CVCs, VCVs or all logatomes for 10 subjects. The error bars show the inter-individual standard

deviations. The model function given in equation (1) was fitted to the data by varying the free parameters SRT (SNR at 55% intelligibility) and s (slope of the psychometric function at the SRT).

$$\Psi(x) = \frac{1-g}{1+\exp(4 \cdot s \cdot (SRT - x))} + g \qquad (1)$$

Here: x: SNR, g: guessing probability (g = 8.9 %) and Ψ: intelligibility. The fit was performed by maximizing the likelihood under the assumption that the recognition of each logatome is a Bernoulli trial (cf. [7]). This yielded a slope of (5.4 ± 0.6) %/dB and a SRT of (-12.2 ± 1.1) dB.

Note that CVCs have always a higher intelligibility than VCVs except for -20 dB SNR.

	d	t	g	k	f	s	b	p	v	ts	m	n	∫	l
d	16	12					8	14	12			10		12
t		62												
g		16	12	24							14			8
k		8	14	26	8							8		
f	8	8			30						10	12		
s						80			18					
b	8	18					10	10			10			8
p					10		12	32						
v			12	10			14	12	8		12			16
ts					28					66				
m	10	10		8			12	8			12			10
n	8	12	12	16				12						16
∫													92	
l	18	12			10	8					8			18

Figure 2: C*onfusion matrix (response rates in %) for normal-hearing subjects at -15 dB SNR, measured with consonants embedded in logatomes. Row: presented phoneme, column: recognized phoneme Grey scales denote different grades of response rates. Response rates below 8% are not shown.*

	a	ɛ	ɪ	ɔ	ʊ	a:	e	i	o	u
a	54		15							
ɛ		79	9		9					
ɪ		18	57							
ɔ				11	24			15	9	18
ʊ				9	24			8	24	14
a:	29					63				
e							84	10		
i							10	78		
o				11	13	8	10	10	25	10
u					11	8			20	41

Figure 3: *Confusion Matrix for normal-hearing subjects at -15 dB SNR, measured with vowels embedded in logatomes. The display is the same as in Fig.2.*

Figure 2 and 3 show the confusion matrices of consonants and vowels for all 10 subjects. Due to the design of OLLO each middle consonant was presented 5 times and every vowel 8 times at a given SNR to each subject. Hence, the overall number of presentations of each phoneme for these matrices are 50 and 80 respectively. The SNR was chosen to -15 dB, which corresponds to an intelligibility of 33% (VCV) and 52% (CVC). Each row symbolizes the presented

phoneme and each column the recognized one. Correct recognized phonemes are shown as diagonal elements of the matrices. Due to clarity all entries below 8% were left blank.

Corresponding to Fig. 2 fricative consonants like "∫", "s" and "ts" are recognized best whereas voiced consonants like "n", "v" and "b" are recognized worst or not at all. Note the big variance between the diagonal elements of "n" and "∫". Unvoiced plosive consonants like "p", "t" and "k" are recognized at a significantly higher recognition rates than voiced ones ("b", "d", "g"). There are almost no confusions between consonants with very high frequency content as "∫", "ts" and those with low one. However there does not seem to be a systematic pattern of confusions.

There is some kind of clustering in the vowel confusion matrix (Fig. 3): "ɔ", "ʊ", "o" and "u" are recognized worst and there are many confusions between them. The next cluster is "a", "a:" with no significant confusions with other vowels. The vowels best recognized are "ɛ", "ɪ", "e" and "i".

3. The perception model

3.1. Specification

The perception model applied in this study was initially developed by Dau et al. [8] and it was further on used to model many different psychoacoustical experiments with different masking conditions as well as modulation detection tasks in an extended version [9]. In this study this extended version is combined with a standard DTW speech recognizer to mimic the decision process in a closed speech intelligibility test.
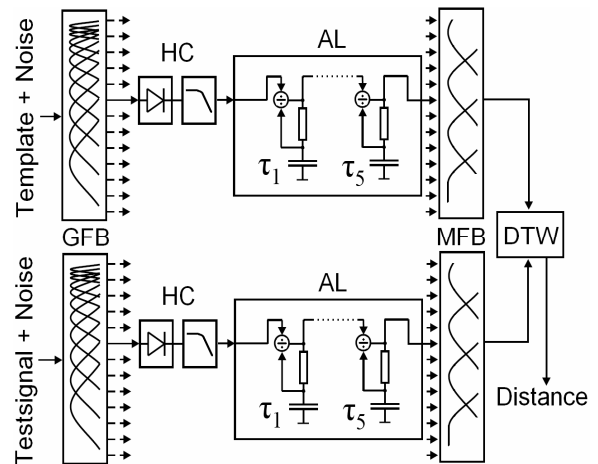

Figure 4: *Scheme of the speech intelligibility model. The model calculates the distance between both, the template waveform and the testsignal waveform after preprocessing in the same way. GFB: gammatone filterbank, HC: hair cell modell, AL: adaptation loops, MFB: modulation-filterbank, DTW: Dynamic-Time-Warp speech recognizer.*

Figure 4 shows the model structure. The level of the template speech waveform is set to 60 dB SPL and both the background ICRA-noise and a hearing threshold simulating noise for normal-hearing listeners is added. The resulting waveform is filtered using a gammatone filterbank ([10]) with 27 frequency channels between 236 Hz and 8 kHz equally spaced on an ERB-scale. The filter-outputs are half-wave rectified and low pass filtered at 1 kHz in a hair cell model.

After processing with five consecutive adaptation loops with time constants chosen as in [3] the signal is again filtered by a modulation filterbank, that consists of 4 modulation filters: one low pass at 2.5 Hz and three band passes with center frequencies of 5, 7.5 and 10 Hz and bandwidths of 5 Hz, respectively. The outcome is an "internal representation" (IR) of the time signal. The testsignal + noise waveform is preprocessed in the same way by the perception model. Note that "noise" in this scheme means running ICRA background noise added to a running hearing threshold simulating noise for normal-hearing subjects. All samples of the training vocabulary were equalized to the same length before processing by attaching silence. This was done to rule out a possible discrimination cue due to the individual length of the speech recordings.

The IR of the template and the IR of the testsignal are the inputs of the speech recognizer, that calculates the Euclidian distance between the two versions. To allow for a mismatch in the temporal structure between sample and template a DTW algorithm [4] performs local stretching and compression of the time axes of both IRs in order to achieve a minimal distance. The logatome with the least distance is chosen as the recognized one. The response alternatives given to the model were the same as for HSR.

Two model configurations were realized in this study:

- In configuration A there were 5 IRs per logatome as templates. None of the 5 original recordings was identical to the tested time signal. The logatome that yielded the minimum mean distance of all 5 IRs was chosen as the recognized one. This mimics a realistic task for common speech recognizers because the exact acoustic utterance is unknown.

- Model configuration B contained a single IR per logatome as a template whereas the original speech material was identical to that of the test signal. Thus the resulting IRs differ only in the initially added background noises. In contrast to configuration A this configuration disregards the natural variability of speech thus it assumes perfect knowledge of the "template" to be matched with the DTW algorithm.

There are many combinations possible to select speech material from OLLO for performing these model calculations. For these two model configurations the speech recognizing task was calculated 10 times using each time a new combination of speech recordings spoken by the same speaker.

3.2. Model predictions and comparison with listening tests

The resulting psychometric functions of the ASR experiments are shown in Fig. 5. Additionally the fitted psychometric function for normal-hearing subjects from Fig. 1 is plotted for a reference. Configuration A shows the same recognition rates for CVCs and for VCVs. The resulting SRT calculated by a fitted psychometric function is 1.3 dB and thus is more than 13 dB higher than that in HSR. It was assumed that in this model configuration, which closely resembles ASR tasks, 100% model recognition rate can never be achieved even without background noise. This is due to the inherent speech variability that is still a problem in ASR tasks [11]. To include this fact a third parameter (the difference between 100% and the saturation recognition rate of the model) was

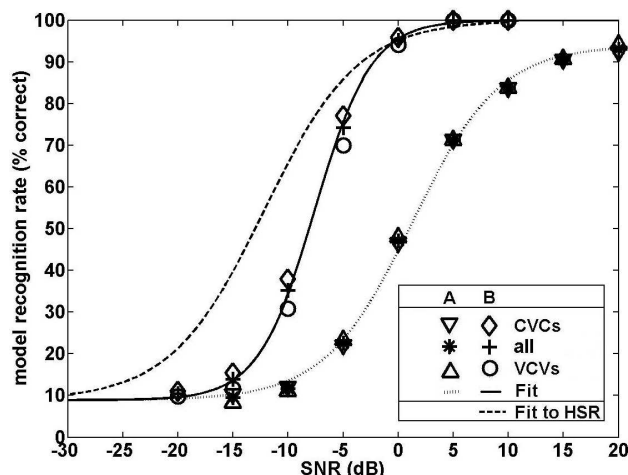introduced into the fitting routine. With a slope of 5.8 %/dB the reference slope is reproduced quite well.

Figure 5: *Predicted psychometric functions for model configurations A and B derived with utterances of logatomes in ICRA-noise at fixed SNR respectively. For comparision inserted: fit to measured normal-hearing psychometric function (HSR) from Fig. 1.*

A much better prediction of the normal-hearing psychometric function is achieved with model configuration B. The order of CVC and VCV as well as the upper part of the reference curve is modelled correctly. 100 % recognition rate is reached at 10 dB SNR. The slope (8.9 %/dB) deviates slightly from the reference, the SRT (-7.6 dB) is much closer to human listeners SRT, but still there is a gap of 4.6 dB between them.

In the following only confusion matrices for model configuration B are evaluated and compared to HSR confusion matrices. The SNR was chosen to -10 dB to ensure about the same intelligibility as for human listeners.

	d	t	g	k	f	s	b	p	v	ts	m	n	∫	l
d	22						8		16		16	20		
t		26					10				22	12		8
g			26				10		8		14	18		8
k			10	20			10				18	16		8
f					12		10		10		18	18		14
s						54					12			10
b							30				16	14		18
p							12	18	12		10	22		10
v	8								24		22	18		8
ts							8		8	38	16	14		
m							10				32	14		18
n							10	8	16		14	26		10
∫												8	72	
l							10		8		18	18		32

Figure 6: *Consonant confusion matrix for model configuration B at -10 dB SNR, The display is the same as in Fig.2.*

Figure 6 and 7 show these confusion matrices. Comparing figure 2 to figure 6 the same consonants "∫", "ts" and "s" are recognized best by the model but that high human recognition rates like 92% for "∫" are not reached. However, some consonants like "n", "v" and "b" are recognized even better

by the model than by human listeners. There is no significant difference between the model recognition rates for unvoiced and voiced plosives. Overall the "contrast" of the model matrix between the diagonal elements is worse than in HSR.

Figure 7: *Vowel confusion matrix for model configuration B at -10 dB SNR, The display is the same as in Fig.2.*

This is also the case for the model confusion matrix for vowels: The clustering found in figure 3 could not be reproduced. At -10 dB SNR the overall recognition rate of CVC utterances is significantly worse than for normal-hearing subjects at -15 dB SNR (38% compared to 52%). However, the phonemes "ɔ" and "ʊ" are recognized slightly better by the model than in HSR. The characteristic nearly uniform columns at "o" and "u" provide an indication that these phonemes are the most probable vowels to recognize by presenting any vowel at that high SNRs.

4. Discussion

Two model configurations were employed, one taking the natural variability of speech into account, the other one disregarding it. Our results show that there is only a chance of predicting the psychometric function for normal-hearing listeners by ignoring the variability of speech itself, i.e. taking identical speech test and training utterances as inputs for the model. Conversely this gives an indication that speech variability is not crucial to speech intelligibility of normal-hearing subjects at high SNRs. Human speech recognition is as perfect and in some phonemes better than the prediction if the listener knew the audio signal before the recognition process. However speech variability is crucial to a model that does not hold the exact speech recording in its training vocabulary.

Although confusion matrices of HSR and ASR are quite similar (especially the consonant phoneme ones), those of the model show a smaller contrast between highly and poorly recognized phonemes. This can be an indication that human listeners use more information from high frequencies to discriminate nonsense speech material than it is done by this model. In each ASR confusion matrix there is a bias favouring some phonemes, like "o" and "u" in Fig. 7, independent of the type of the presented phoneme. This bias could be corrected by changing the selection criteria, which would probably also be done by human listeners during the measurement procedure.

5. Conclusions

This speech intelligibility model is based on the time signal of speech and consists of a psychoacoustically motivated preprocessing and a simple speech recognizer. It is capable of predicting essential aspects of speech intelligibility of normal-hearing subjects. By considering the intrinsic variability of speech the modeled SRT is 13 dB higher than human listeners show. This is consistent with findings of other studies exploring the human-machine gap [1]. Introducing a perfect knowledge about the speech signal to recognize allows for predicting the psychometric function with a much smaller offset. This refers to the "optimal detector" concept required to model human perception assuming that the "world knowledge" yields an optimal template in each HSR experiment. In addition it was possible to detect characteristic differences between phoneme confusion matrices of HSR and ASR.

Future studies should investigate speech intelligibility of hearing-impaired subjects and also should analyse the influence of the loss of dynamic range at the hearing-impaired on speech intelligibility in a microscopic way.

6. Acknowledgements

7. References

[1] Meyer, B., T. Brand, and B. Kollmeier, *Phoneme Confusions in Human and Automatic Speech Recognition,* this issue.

[2] ANSI, *ANSI S3.5-1997 - Methods for Calculation of the Speech Intelligibility Index.* 1997.

[3] Holube, I. and B. Kollmeier, *Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model.* J. Acoust. Soc. Am., 1996. **100**(3): p. 1703-16.

[4] Sakoe, H. and S. Chiba, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978. **ASSP-26**(1): p. 43-49.

[5] Wesker, T., et al., *Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines*, in *Interspeech 2005,* p. 1273-1276.

[6] Dreschler, W.A., et al., *ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment.* Audiology, 2001. **40**: p. 148–157.

[7] Brand, T. and B. Kollmeier, *Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests.* J. Acoust. Soc. Am., 2002. **111**(6): p. 2801-2810.

[8] Dau, T., D. Püschel, and A. Kohlrausch, *A quantitative model of the "effective" signal processing in the auditory system: I. Model structure.* J. Acoust. Soc. Am., 1996. **99**: p. 3615-3622.

[9] Dau, T., *Modeling auditory processing of amplitude modulation.* J. Acoust. Soc. Am., 1997. **101**: p. 3061(A).

[10] Hohmann, V., *Frequency analysis and synthesis using a Gammatone filterbank.* Acta acustica / Acustica, 2002. **88**(3): p. 433-442.

[11] Lippmann, R.P., *Speech recognition by machines and humans.* Speech Communication, 1997. **22**(1): p. 1-15.