The Interspeech 2008 Consonant Challenge

Martin Cooke [1], *Odette Scharenborg* [2]

[1] Department of Computer Science, University of Sheffield, UK
[2] Centre for Language and Speech Technology, Radboud University, Nijmegen, The Netherlands
m.cooke@dcs.shef.ac.uk, o.scharenborg@ru.let.nl

Abstract

Listeners outperform automatic speech recognition systems at every level, including the very basic level of consonant identification. What is not clear is where the human advantage originates. Does the fault lie in the acoustic representations of speech or in the recognizer architecture, or in a lack of compatibility between the two? Many insights can be gained by carrying out a detailed human-machine comparison. The purpose of the Interspeech 2008 Consonant Challenge is to promote focused comparisons on a task involving intervocalic consonant identification in noise, with all participants using the same training and test data. This paper describes the Challenge, listener results and baseline ASR performance.

Index Terms: consonant perception, VCV, human-machine performance comparisons

1. Motivation

In most comparisons of human and machine performance on speech tasks, listeners win [1][2][3] (but see [4]; for an overview, see0[5]). While some of the benefit comes from the use of high-level linguistic information and world knowledge, listeners are also capable of better performance on low-level tasks such as consonant identification which do not benefit from lexical, syntactic, semantic and pragmatic knowledge. This is especially the case when noise is present. For this reason, understanding consonant perception in quiet and noisy conditions is an important scientific goal with immediate applications in speech perception (e.g. for the design of hearing prostheses) and spoken language processing [6]. A detailed examination of confusion patterns made by humans and computers can point towards potential problems at the level of speech signal representations or recognition architectures. For example, one compelling finding from a number of recent studies has been that much of the benefit enjoyed by listeners comes from better perception of voicing distinctions [7][8][9].

A number of corpora suitable for speech perception testing exist [7][10], although few contain sufficient data to allow training of automatic speech recognizers. However, the main motivation for the Interspeech 2008 Consonant Challenge was not solely to make available a corpus large enough for human-machine comparisons, but also to define a number of varied and challenging test conditions designed to exercise listeners and algorithmic approaches. In addition, by providing software for perceptual testing and scoring, the aim was to support a wide range of comparisons, for both native and non-native listeners.

This paper describes the design, collection and post-processing of the Consonant Challenge corpus and specifies the test conditions as well as the training and development material. It provides results for native listeners and for two baseline automatic speech recognition systems.

2. Corpus

2.1. Design

The corpus consists of intervocalic English consonants (VCV), for a number of vowel and stress combinations. The 24 consonants (/b/, /d/, /g/, /p/, /t/, /k/, /s/, /sh/, /f/, /v/, /dh/, /th/, /ch/, /z/, /zh/, /h/, /dj/, /m/, /n/, /ng/, /w/, /r/, /y/, /l/) were combined with nine vowel contexts consisting of all possible combinations of the three vowels /i:/ (as in "beat"), /u:/ (as in "boot"), and /ae/ (as in "bat"). Each VCV was produced using both initial and final stress (e.g. 'aba versus ab'a) leading to a total of 28 (speakers) * 24 (consonants) * 2 (stress types) * 9 (vowel contexts) = 12,096 tokens.

2.2. Speakers

Twelve female and 16 male native English speakers aged between 18-49 contributed to the corpus. Speakers originated from various regions of the UK, although most were born within 50 km of Sheffield. None had a strong regional accent.

2.3. Recording

Recordings were made in an IAC single-walled acoustically isolated booth at the University of Sheffield. Speech material was collected from a single Bruel & Kjaer (B & K) type 4190 ½ -in. microphone placed 30 cm in front of the talker. The signal was pre-amplified by a B & K Nexus model 2690 conditioning amplifier prior to digitisation at 50 kHz by a Tucker-Davis Technologies System 3 RP2.1.

Speakers produced VCVs in isolation by reading out tokens presented on a computer screen, and were given both verbal and written instructions on how to interpret token names, with a particular focus on /th/, /dh/, /dj/, and /zh/. Speakers ran through a practice with the experimenter before speaking alone in the booth. Speakers produced all VCVs with initial stress, followed by final stress. Collection of speech material was under computer control. Although VCV tokens are not "normal", speakers were asked to produce them at a "normal" speaking rate, to avoid problems with lengthy, drawn-out productions sometimes found in VCV corpora.

2.4. Post-processing

Signals were high-pass filtered at 50 Hz to remove low frequency noise, endpointed, downsampled to 25 kHz and normalized to have the same RMS level. Tokens were screened to check for poor or mispronunciations, endpointing problems or extraneous noise. This led to the identification of 301 unusable tokens (2.5% of the corpus), of which 16% were irrecoverable endpointing errors and 4% contained noise. Rejection of the remaining 80% was due to pronunciation problems, mostly for the consonants /th/, /dh/ and /zh/.

Screening uncovered several other phenomena: /ng/ was sometimes realized as /n/+/g/; complete vowel reduction was occasionally observed, principally for /ae/; there was some centralization of /i:/ and /u:/; and frequent incorrect stress assignment. These tokens were retained in the corpus.

test set	noise type	SNR (dB)
1	clean	
2	competing talker	-6
3	8-speaker babble	-2
4	speech-shaped noise (SSN)	-6
5	factory noise	0
6	modulated SSN	-6
7	3-speaker babble	-3

Table 1. Test sets for the Consonant Challenge

2.5. Test, development and training sets

To facilitate comparisons of human and machine performance using identical material, subsets of the corpus for testing, training and development purposes were specified. Different speakers were used for each of the three subsets.

Seven test sets were produced to accommodate clean and 6 noise conditions, using material from 4 male and 4 female talkers. Each test set contains 2 instances of each of the 24 consonants from each speaker (i.e. 384 tokens per test set). Table 1 summarises the seven test conditions. The 6 noise types were chosen to provide a varied range of challenging conditions, which, with the exception of SSN, are nonstationary. All noises apart from factory noise have a long-term spectrum equivalent to that of speech. Many noises can be expected to induce informational [11] as well as energetic masking. In particular, 8-talker babble has been shown to be a particularly effective informational masker of VCV tokens [12]. Modulated SSN (test set 6) was produced by multiplying a speech-shaped noise signal with the short-term envelope of sentence material. Modulated SSN introduces the temporal fluctuations of a speech masker but is not intelligible, so has little or no informational masking effect.

Signal-to-noise ratios were chosen via pilot tests to produce similar overall identification scores in the range 65-75% in each condition (in the event, the range 66-79% was obtained). The aim was to avoid ceiling performance for listeners and floor effects for algorithms. Tokens were added to noise samples of duration 1.2s. Rather than co-gating the VCV and noise, the onset time of the VCV relative to the noise was varied in order to make the appearance of the VCV unpredictable in the noise. Onsets took one of 8 values linearly-spaced in the range 0 to 400 ms. Each consonant occurred the same number of times at each of the 8 onsets. For each token, the noise signal was scaled to produce the required SNR in the region where the speech was present.

For the noisy conditions, test material for the Challenge is available in both single-channel and dual-channel versions. The former contains the mixed VCV plus noise, while the latter provides the VCV and noise on separate channels. The dual-channel versions are provided to allow the evaluation of models which make assumptions about some stages of human consonant perception, or to allow the estimation of "ideal" performance ceilings for algorithms.

After the removal of unusable tokens, a training set of 6664 clean tokens was created using material from 8 male and 8 female speakers. Seven development sets consisting of 192 tokens each (2 of each consonant from 4 male speakers) was produced using the same noise types as used in the test sets.

3. Human consonant identification

3.1. Listening tests

Twenty seven native English listeners aged between 18 and 48 who reported no hearing problems identified the 384 VCVs of the test set. Listeners were drawn from the staff and students at the University of Sheffield and were paid for their participation. Perception tests ran under computer control in the IAC booth. Listeners were presented with a screen layout (Figure 1) in which the 24 consonants were represented using both ASCII symbols and with an example word containing the sound. Listeners were phonetically-naive and were given instructions as to the meaning of each symbol. They underwent a short practice session prior to the main test. Two listeners failed to reach a criterion level of 85% in a practice session using clean tokens. Another failed to complete all conditions, while a fourth was an outlier on most of the test conditions. Results are reported for the remaining 23 listeners. For the main test, listeners started with the clean condition. The order of the noisy conditions was randomised.

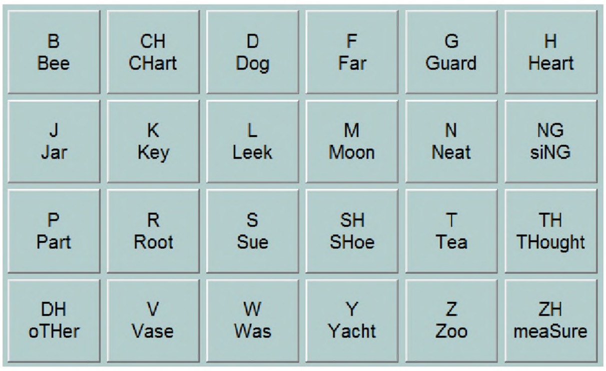

Figure 1. Screen layout for perception tests.

Test set	1	2	3	4	5	6	7
Rec. rate	93.8	79.5	76.5	72.2	66.7	79.2	71.4
Std. err.	0.57	0.78	0.79	0.75	0.77	0.61	0.74

Table 2. Native listener scores.

3.2. Results

3.2.1. Overall identification scores

Recognition rates and standard errors are shown in Table 2. Note that comparisons across some subsets of noise conditions are not meaningful since noises were mixed at a variety of SNRs. However, comparisons between those noises presented at an SNR of -6 dB confirm trends found in previous studies. For instance, a competing talker is a significantly weaker masker than stationary speech-shaped noise [13] and, for this task, performance was indistinguishable for competing speech and noise modulated by speech, a result which suggests that informational masking is not a major factor for the competing speech masker on a VCV task, confirming [12]. The factory noise background, presented at the least severe SNR, proved the most challenging type of noise. Comparison of its long-term average spectrum with that of speech suggests that, when both are normalized to have the same overall energy, factory noise has less energy than speech in the region below 800 Hz but substantially more in the 800-3500 kHz region where perceptually-important speech information lies.

3.2.2. Consonant error rates

Figure 2 depicts error rates for individual consonants in each test condition. For the quiet condition, a small group of consonants accounted for most of the errors. Listeners had particular problems with the dental fricatives, /dh, th/, the labiodentals fricatives /f, v/ and the palatals /zh/ and /dj/. Some of these sounds were also responsible for a large number of production errors, suggesting that poor orthographic-phonemic correspondence during production was part of the problem, although the relatively low error rate for /dj/ and /zh/ in noise suggests that orthography was not a limiting factor for these sounds.

The second panel of figure 2 shows the mean error rate per consonant across all noisy conditions. As a group, the sibilants were typically well identified while /f, v, th, dh, b, ng/ presented most difficulties. These rankings are very similar to those found in a recent study which employed stationary speech-shaped noise [14].

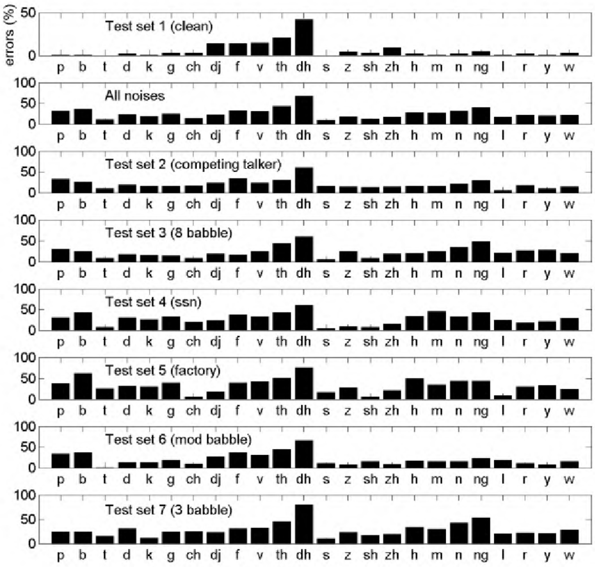

Figure 2. Averaged listener error rates per consonant.

3.2.3. *Confusions*

Confusion matrices for the clean test set and averaged over the noisy conditions are shown in figure 3. In both quiet and noise, most confusions occur within the set /f, v, th, dh/. In quiet, /dj/ and /zh/ are frequently confused, perhaps due to symbol confusion, while in both quiet and noise /ng/ is sometimes heard as /g/, probably reflecting incorrect realisations. In noise, /b/ and /v/ are often confused, as found elsewhere [15]. /v/ is the most reported sound in noise (1.56 times its actual rate of occurrence), followed by /g/ (1.33), while /dh/ (0.74) and /ng/ (0.81) are least reported.

3.2.4. *Transmitted information analysis*

A standard way to summarise perceptual confusions since Miller and Nicely [15] is to measure the proportion of transmitted information (TI) for consonantal features. Figure 4 shows TI measures for manner, place and voicing. For quiet, voicing is least well transmitted (that is, listeners report a voiced sound when an unvoiced sound was present, and vice versa), largely due to confusions amongst the dental and labiodental fricatives. However, averaged over most of the noisy test sets, the three features are equally-well transmitted. The TI analysis suggests that place confusions, perhaps based on spectral masking, are in the main responsible for the difficulty listeners had with factory noise. Voicing information is particularly adversely affected by stationary noise. Place is less confused for modulated speech shaped noise than for competing speech while for manner and voicing the opposite is true.

HEARD

	p	b	t	d	k	g	ch	dj	f	v	th	dh	s	z	sh	zh	h	m	n	ng	l	r	y	w
p	99	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
b	1	99	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
d	.	.	.	98	.	2	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
k	.	.	.	.	99	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
g	.	.	.	.	1	97	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.
ch	.	.	2	.	.	.	97	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
dj	.	.	.	2	.	4	2	86	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	.
f	.	.	.	.	.	.	.	.	86	1	11	3	.	.	.	.	.	.	.	.	.	.	.	.
v	.	.	.	.	.	.	.	.	4	85	3	7	.	.	.	.	1	.	.	.	.	.	1	.
th	.	.	.	.	.	.	.	.	6	.	79	11	4	.	.	.	.	.	.	.	.	.	.	.
dh	.	.	.	.	.	.	.	.	.	17	23	58	.	1	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.
z	.	.	.	.	.	.	.	.	.	.	.	.	1	96	.	2	.	.	.	.	.	.	1	.
sh	.	.	.	.	.	.	1	.	.	.	.	.	1	.	97	1	.	.	.	.	.	.	.	.
zh	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	91	.	.	.	.	.	.	.	.
h	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	98	.	.	.	.	.	1	1
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	99	.	.	.	.	.	.
n	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	98	.	2	.	.	.
ng	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	1	95	.	.	.	.
l	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	99	.	.	.
r	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	98	.	1
y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	99	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	97

HEARD

	p	b	t	d	k	g	ch	dj	f	v	th	dh	s	z	sh	zh	h	m	n	ng	l	r	y	w
p	69	6	2	1	6	3	.	.	2	3	2	1	1	.	.	.	3	1	.	1	1	1	.	1
b	5	64	.	1	1	3	.	.	1	14	1	2	.	.	.	.	1	1	.	1	1	1	.	2
t	1	.	89	1	4	.	1	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.
d	.	1	4	76	.	6	.	1	.	1	1	3	1	1	.	.	.	.	2	1	1	1	1	.
k	4	1	2	.	82	7	.	.	.	1	.	.	.	.	.	.	2	.	.	1	1	.	.	.
g	1	2	1	4	6	76	.	.	.	1	.	.	.	.	.	.	1	.	.	7	.	.	1	1
ch	.	.	3	.	1	.	86	7	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.
dj	.	.	2	1	1	5	6	78	.	.	.	.	.	.	.	6	.	.	.	1	.	.	.	.
f	2	1	.	.	.	.	.	.	68	7	16	3	1	.	.	.	1	.	.	.	.	.	.	.
v	2	8	.	1	1	1	.	.	3	69	3	8	.	.	.	.	.	.	.	.	.	2	.	1
th	1	1	2	.	1	.	.	.	17	3	57	13	3	1	.	.	.	.	.	.	.	.	.	.
dh	1	5	1	6	.	2	.	.	1	29	12	33	1	4	.	1	.	.	1	1	2	1	.	.
s	.	.	.	.	.	.	.	.	2	.	3	1	90	3	.	.	.	.	.	.	.	.	.	.
z	.	1	.	1	.	.	.	.	.	5	1	5	2	82	.	2	.	.	.	.	.	.	.	.
sh	.	.	.	.	.	.	2	1	.	.	.	.	1	.	89	5	1	.	.	.	.	.	.	.
zh	.	.	.	.	.	1	.	9	.	1	.	1	.	1	1	84	1	.	.	.	.	1	1	.
h	3	3	.	1	2	2	.	.	1	4	2	1	.	.	.	.	72	.	1	1	1	3	2	2
m	1	4	.	.	.	.	.	.	.	4	.	1	.	.	.	.	.	73	6	1	3	4	.	3
n	1	1	1	2	.	1	.	.	.	1	1	1	.	.	.	.	.	9	68	2	8	3	1	1
ng	.	1	1	2	1	19	.	.	.	2	.	.	1	.	.	.	1	2	5	60	1	2	1	1
l	.	2	.	1	.	1	.	.	.	2	.	1	.	.	.	.	.	1	2	.	84	3	1	2
r	.	2	.	.	.	1	.	.	.	5	.	.	.	.	.	.	.	2	1	1	3	79	1	6
y	.	.	.	1	.	4	.	.	.	1	.	.	1	1	.	.	1	.	1	3	3	1	80	2
w	.	3	.	.	1	1	.	.	.	3	.	.	.	.	.	.	1	1	1	.	2	7	2	78

Figure 3. Confusion in for clean (top) and averaged over the noise conditions (bottom) expressed as percentages. Rows represent sounds 'sent' and columns 'heard'.

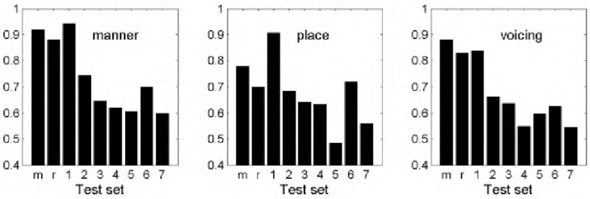

Figure 4. Proportion information transmitted for manner, place and voicing. Results for the baseline recognisers are also shown (m=MFCC,r=ratemap).

4. Baseline recognizer

4.1. Recognizer structure and training

The performance of various acoustic features and recognizer architectures (monophone, triphone, gender-dependent/ independent) was investigated. Two representative combinations were chosen as baselines for the Consonant Challenge. The best performance on the clean test set (88.5%) was obtained using a 24-mixture CDHMM system with 3 state monophone models, based on the "standard" 39-dimensional MFCC_0_Z_D_A feature. Separate HMMs for initial and final vowels were used. Ratemaps, an auditory filterbank-based representation (see [8] for more details), was chosen as an alternative acoustic feature vector. These achieved a score of 84.4% using 64-dim feature vectors and the same model architecture. HMMs were trained from a flat start using HTK [17].

4.2. Results

Figure 5 compares consonant error rates for listeners and the two baseline recognisers on the clean test set. Listener-machine errors are strongly-correlated for MFCCs (r=0.81, p < 0.0001) and less so for ratemaps (r=0.54, p < 0.01), with both humans and machines having most difficulty with the dental fricatives. Differences in human and machine performance are highest for the plosives (apart from /b, p/) and the nasals.

The errors for /d/ and /g/ for MFCCs are due to manner confusions (with /dj/ and /ng/, respectively). As with humans, /d/ is also confused with /g/. The errors for ratemaps for /d/ are due to confusions with /dj/ (manner), while the /g/ errors are place related (confusions with /b/). For the nasals, the confusions for the MFCC system are mainly within the nasal class, but for ratemaps /ng/ is most often confused with /l/ (manner+place confusion). While humans confuse /ng/ most often with /g/, machines seem to have fewer problems with /ng/. However, in interpreting confusions it should be noted that listener scores are averaged over 23 listeners, whereas the baseline systems are equivalent to a single ‘listener’.

A transmitted information analysis (figure 4, column ‘m’) showed that the MFCC baseline outperformed listeners on the voicing feature but was significantly worse for place. As expected on the basis of the overall recognition performance, ratemaps have the lowest transmitted information scores (figure 4, column ‘r’), although the performance for the voicing feature is only slightly lower than for humans.

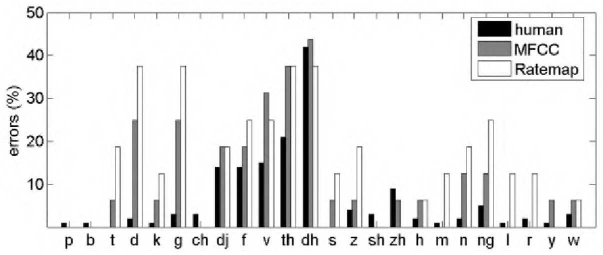

Figure 5. Human and baseline ASR error rates.

5. Concluding remarks

The Interspeech 2008 Consonant Challenge aims to promote human-machine, machine-machine and human-human (native/non-native) comparisons on a consonant identification task which avoids the use of high-level speech knowledge. A VCV corpus was collected consisting of a number of varied and challenging test conditions specifically designed for performance comparisons. Listener and baseline recogniser results are reported. For certain features such as voicing, an MFCC-based HMM baseline outperformed listeners in the noise-free condition, but fell far short of listeners for the place feature. Understanding the basis for these differences is a goal for future research.

All materials associated with the Consonant Challenge can be accessed at [18].

6. Acknowledgements

The authors would like to thank María Luisa García Lecumberri, Youyi Lu, Ning Ma, Vincent Wan and Matt Gibson for helpful discussions. This work was supported by a grant from the EU PASCAL Network of Excellence, the EU Marie Curie Research Training Network “Sound to Sense” and a Veni-grant from the Netherlands Organisation for Scientific Research (NWO) to Odette Scharenborg.

7. References

[1] Lippmann, R., “Speech recognition by machines and humans”, Speech Comm., 22 (1), 1997, 1-15.

[2] Barker, J., Cooke, M. P., “Modelling speaker intelligibility in noise”, Speech Comm., 49, 2007, 402-417.

[3] Alwan, A., Zhu, Q., Lo, J. “Human and machine recognition of speech sounds in noise”, Conf. Sys., Cybernetics, and Information, XIII, 2001, 218-223.

[4] Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., “Super-Human Multi-Talker Speech Recognition: The IBM 2006 Speech Separation Challenge System”, Interspeech, Pittsburgh, PA, 2006.

[5] Scharenborg, O., “Reaching over the gap: A review of efforts to link human and automatic speech recognition research”, Speech Communication, 49, 2007, 336-347.

[6] Holube, I., Kollmeier, B. “Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model,” J. Acoust. Soc. Am., 100, 1996, 1703-1716.

[7] Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B., “A human-machine comparison in speech recognition based on a logatome corpus”, Proc. Speech Recog. and Intrinsic Variation, Toulouse, France, 2006.

[8] Cooke, M., “A glimpsing model of speech recognition in noise”, J. Acoust. Soc. Am., 119 (3), 2006, 1562-1573.

[9] Sroka, J., Braida, L., “Human and machine consonant recognition”, Speech Comm., 45, 2005, 401-423.

[10] Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X., “Consonant recordings for speech testing”, J. Acoust. Soc. Am., 106, 1999, L71-L74.

[11] Carhart, R., Tillman, T. W. Greetis, E. S., “Perceptual masking in multiple sound backgrounds”, J. Acoust. Soc. Am., 45, 1969, 694-703.

[12] Simpson, S., Cooke, M. P., “Consonant identification in N-talker babble is a nonmonotonic function of N”, J. Acoust. Soc. Am., 118, 2005, 2775-2778.

[13] Bronkhorst, A. W. Plomp, R., “Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing,” J. Acoust. Soc. Am., 92, 1992, 3132-3139.

[14] Phatak, S.A., Allen, J.B., “Consonant and vowel confusions in speech-weighted noise”, J. Acoust. Soc. Am., 121, 2007, 2312-2326.

[15] Garcia Lecumberri, M.L. and Cooke, M. P. (2006) Effect of masker type on native and non-native consonant perception in noise, J. Acoust. Soc. Am. *119, 2445-2454*

[16] Miller, G. A., Nicely, P., “An Analysis of Perceptual Confusions among some English Consonants”, J. Acoust. Soc. Am., 27, 1955, 338-352.

[17] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK book (for HTK version 3.2). Techn. Report, Cambridge University, Eng. Dept., 2002.

[18] http://www.odettes.dds.nl/challenge_IS08/