7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Memory of AMR coded speech distorted by packet loss

Arne Nykänen
Luleå University of Technology, 971 87 Luleå, Sweden.

David Lindegren
LM Ericsson AB - Ericsson Research, 97753 Luleå, Sweden.

Louisa Wruck
Luleå University of Technology, 971 87 Luleå, Sweden.

Robert Ljung
University of Gävle, 801 76 Gävle, Sweden.

Johan Odelius
Luleå University of Technology, 971 87 Luleå, Sweden

Sebastian Möller
Quality and Usability Lab, Telekom Innovation Labs, TU Berlin, 10587 Berlin, Germany

Summary
Previous studies have shown that free recall of spoken word lists is impaired if the speech is presented in background noise, even if the signal-to-noise ratio is kept at a level allowing full word identification. The objective of this study was to examine recall rates for word lists presented in noise and word lists coded by an AMR (Adaptive Multi Rate) telephone codec distorted by packet loss. Twenty subjects performed a word recall test. Word lists consisting of ten words were played to the subjects. The subjects repeated each word immediately after it had been played, to ensure that the words were heard correctly. After the complete list had been played the subjects wrote down all words remembered. In this way, both word identification and recall rates were measured. Three distorted conditions were compared with an undistorted control condition using a within-subject design: speech spectrum weighted noise at 4 dB SNR, and AMR coded speech with two levels of packet loss, one mild and one severe. The results confirmed the disruptive effect of noise on free recall of words, while no significant impairment was found for the AMR distortions. The noise and the AMR coding with mild packet loss gave approximately the same impairment of word identification. The AMR coding with severe packet loss gave a larger impairment of word identification, even though the word recall rate was unaffected. This result suggests that packet loss in AMR coded speech causes distortions which disrupt recall of words less than noise at levels resulting in the same change of word identification rates. Since impairment of word identification rates did not correlate with impairment of word recall rates models for quality prediction of speech reproductions should not be based on identification rates alone.

PACS no. 43.71.+m, 43.72.+q

1. Introduction

There is by now considerable evidence that the recall of spoken messages is impaired by poor listening conditions, even when people are able to identify what is said correctly [1], [2], [3], [4]. This has been shown for free recall of word lists presented in noise [5] and with long reverberation time [6], and it has been shown for cued recall of spoken lectures under similar impoverished listening conditions [1]. The theoretical explanation for this is cognitive load; the processing of speech in poor listening conditions requires the listener to consider more alternative interpretations and therefore understanding relies more on stored information than in the case under good listening conditions. The perceptual coding

of the speech, which in good listening conditions is largely automatic, becomes more of a controlled resource demanding process. As a consequence, higher demands are placed on limited cognitive processing resources, thereby impairing the encoding and consolidation processes required for the mnemonic retention of the to-be-remembered information. In support of this theory, Ljung et al. [7] provided evidence that low signal quality affects cognitive processes much earlier than it affects speech intelligibility thresholds. The hypothetical functions in Figure 1 show that the memory function declines earlier, and is steeper than the function for speech intelligibility. Thus, demanding listening conditions result in the allocation of more cognitive resources to speech perception, at the expense of cognitive encoding processes. One way in which this relationship can be explored is to examine memory tasks whereby the perceptual degradedness of the to-be-remembered information is varied.

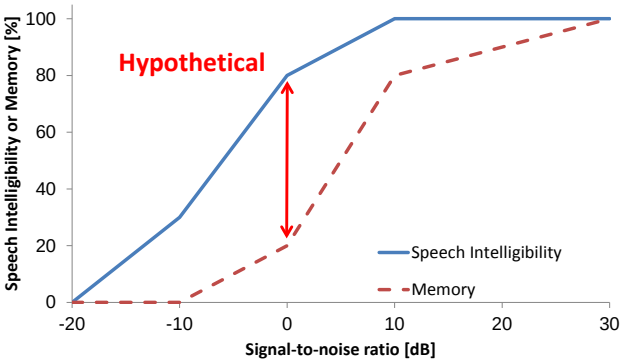

Figure 1. Hypthetical speech intelligibility and memory functions for varying signal-to-noise ratios for speech.

As more telecommunication is moved from analog to digital networks (e.g. GSM and IP-telephony), new effects of degraded speech become apparent. Most current research focus on speech quality and so-called Quality of Experience [8], [9]. Standards for speech quality assessment have been set up by the International Telecommunication Union (ITU). Typically speech quality is assessed by the Mean Opinion Score (MOS) obtained in listening tests (ITU-T P.800). Prediction models, e.g. POLQA (ITU-T P.863) and the E-model (ITU-T G.107), are efficient and accurate in predicting MOS for known kinds of transmission errors. However, they estimate MOS as an indicator of perceived speech quality. There is a distinct lack of research on how speech comprehension is affected by digital telephony. Sheffield et al. [10] showed that memory was degraded for low bit rate digital radio coders (9 to 24 kbps). These coders use similar techniques and bit rates as that typically used in telephone systems. They also showed that memory of speech is not necessarily correlated with perceived speech quality. Further, the extensive knowledge on the detrimental effect of noise on speech – and how it is affected by age and hearing loss – is almost exclusively described as a ratio between speech and noise level. Analog distortions are easily explained in the same discourse whereas digital errors are not. This emphasizes the importance of studying how speech comprehension is affected by digital distortions.

The aim of this study was to assess how speech comprehension is affected by AMR coding and IP packet loss occurring in digital telephony. Memory for speech was used as an indicator for cognitive load from the processing of speech. The objectives were to measure how errors and artifacts caused by packet loss affect memory for speech for normal hearing participants, and to compare this with effects of speech spectrum weighted noise on memory of speech.

2. Method

Methods developed for the assessment of the degradation of memory of speech caused by noise and long reverberation times were used [1], [5].

2.1. Participants

In total, 20 subjects participated in the study (12 female and 8 male with age ranging from 18 to 35 years and mean age 25.4 years). All participants were native speakers of Swedish and all had self-reported normal hearing.

2.2. Experimental design and stimuli

A within-subject design with four conditions was used: 1. Control condition: undistorted speech, 2. Speech distorted with speech spectrum weighted noise at 4 dB SNR, 3. AMR coded speech at 12.2 kbit/s with a low packet loss rate, and 4. AMR coded speech at 4.75 kbit/s with a high packet loss rate. The speech stimuli were taken from a standardized test for speech audiometry [11] played at an equivalent sound level of 64 dBA. Originally, the word lists consist of 40 phonetically balanced one or two syllable words read by a male speaker. For this experiment 16 lists consisting of 10 words each were created by cutting out 10 word segments from the original material, see Table I. In between each word a 5 s

Table I. Word lists compiled from Hagerman's material for speech audiometry tests [11].

1	vas	fisk	frukt	hål	nerv	kalv	näbb	snår	blad	hiss
2	sill	rad	blixt	sol	plats	torsk	dans	stork	kök	spalt
3	doft	skärp	lik	grabb	stund	park	slips	bly	mjölk	frö
4	puss	skydd	grav	damm	dräng	mur	glass	kust	dvärg	sits
5	eld	gift	köp	slant	skämt	lån	chef	bär	häl	pil
6	släkt	ring	spis	sko	fru	kam	kniv	haj	brev	skur
7	stänk	kött	fjäll	spår	blink	tand	död	knä	tjur	skratt
8	fat	folk	glas	hatt	barr	broms	rum	knapp	svan	bänk
9	storm	djur	burk	fluga	zoo	yta	hud	kjol	vete	groda
10	yxa	vinge	palm	byxa	kyrka	mus	bäck	lax	post	nymf
11	regn	bonde	gräs	skrik	löv	stuga	ben	skrift	land	ägg
12	virus	matta	eka	blus	polis	dass	hals	mössa	fält	liv
13	kung	gran	mynt	vante	natt	dam	bild	tumme	gevär	kropp
14	strand	frack	tak	duk	plåt	sax	gös	stock	tall	hjul
15	keps	päron	berg	vev	dröm	häst	sked	bil	lugg	svala
16	puls	klass	ask	lön	lärka	pedal	hall	skåp	kliv	vik

pause was inserted. The experiment was conducted in a sound insulated recording studio using headphones (Head Acoustics HPS IV).

2.2.1. Specification of conditions

Condition 1 was a control condition consisting of undistorted speech recorded with 44.1 kHz sampling frequency and 16 bit dynamic resolution (background noise gave 30 dB SNR). Condition 2 consisted of the speech from Condition 1 distorted with speech spectrum weighted noise at 4 dB SNR. The noise was created by filtering white noise through a second order Butterworth bandpass filter with cut-off frequencies of 125 and 500 Hz, see Figure 2.

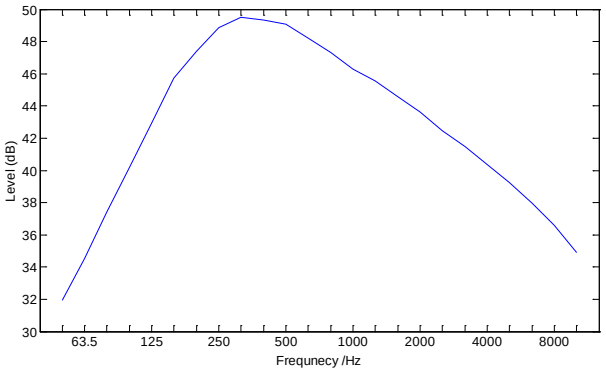

Figure 2. Energy distribution of the speech spectrum weighted noise used in Condition 2.

Conditions 3 and 4 consisted of speech transmitted through a simulated network using AMR-NB coding. The transmission rates were 12.2 kbit/s in Condition 3 and 4.75 kbit/s in Condition 4. The coded speech signal was packetized with one speech frame representing 20 ms of speech per packet [12]. The packet stream was then subject to packet losses introduced during active speech to degrade each word. To ensure that the distortions introduced by packet loss did not destroy segments long enough to affect speech intelligibility, no more than 3 consecutive packets could be lost and no more than 5 packets per word in total. This was also double checked manually to ensure good speech intelligibility. In Condition 3 a packet loss rate of 0.5 packets/s was used and in Condition 4 the packet loss rate was 0.8 packets/s, note that all those packets where lost during an active period of speech and not during silence.

2.2.2. Memory task

16 word lists, each consisting of 10 words were presented to the subjects, four for each condition (1. Control, 2. Noise, 3. Low packet loss rate and 4. High packet loss rate). The task was to memorize the words for later recall. There was a 5 s pause between each word. The subject repeated the word aloud during this pause. This allowed the experiment leader present in the room to check whether the subject had identified the word correctly. In Condition 2 the noise was played continuously (also in the pauses) during the presentation of the list. The conditions were randomly assigned to a wordlist (4 word lists for each condition) and the presentation order of the word lists was randomized for each subject. When one list was finished the subject was asked to write down all words remembered using a computer interface. There was no time limit for recalling and writing down the remembered words. Recall performance was measured for each list as the ratio between the number of correctly recalled

words and the number of correctly identified words.

3. Results

3.1. Intelligibility

The aim of the study was to measure word recall rates for speech presented with a quality allowing full word identification. Therefore, word identification rates were first measured and compared to the undistorted control condition. As the subjects were asked to repeat each word immediately after being presented, the word identification rate could be measured as the proportion of correctly repeated words. Word identification rates on group level are presented in Table II.

Table II. Word identification rates under the four studied conditions.

Cond. 1 (Control)	Cond. 2 (Noise)	Cond. 3 (Low packet loss rate)	Cond. 4 (High packet loss rate)
99.5%	95.0%	95.3%	90.4%

Within-subject differences in word identification rates between the conditions were analyzed using one way repeated measures ANOVA. Due to loss of data, only the last 11 subjects are included in this analysis. Mauchly's test of sphericity showed that the assumption of sphericity was not violated ($\chi^2(5)=6.822$; $p=.237$). The ANOVA showed that there were significant differences in word identification rates between the conditions ($F(3, 30)=18.18$; $p<.05$). In order to compare the conditions without making too many pairwise comparisons three a priori contrasts were selected: Condition 1 (control) vs. Condition 2 (noise), Condition 1 (control) vs. Condition 3 (low packet loss rate) and Condition 1 (control) vs. Condition 4 (high packet loss rate). Results from pairwise t-tests of these a priori contrasts are shown in Table III.

The results show that speech intelligibility measured as proportion of correctly identified words decreased when the distortions were introduced. Condition 2 (noise) and Condition 3 (AMR with low packet loss rate) were similar. Condition 4 (AMR with high packet loss rate) resulted in worse speech intelligibility.

Table III. Within-subject differences in proportions of correctly identified words, tests of a priori contrasts.

Contrast	Mean Diff.	t(10)	Sig. (2-tailed)	r
Control - Noise	5.2%	5.68	.000*	.87
Control - Low packet loss rate	4.8%	4.61	.001*	.82
Control - High packet loss rate	10.7%	6.46	.000*	.89

*Significant difference. Bonferroni adjustment for multiple comparisons requires $p\leq.017$ for a difference to be significant at 95% confidence level.

3.2. Recall of spoken words

The recall rate was measured as the ratio between the number of correctly recalled words and the number of correctly identified words (see Section 3.1.). Word recall rates on group level are presented in Table IV.

Table IV. Word recall rates under the four studied conditions.

Cond. 1 (Control)	Cond. 2 (Noise)	Cond. 3 (Low packet loss rate)	Cond. 4 (High packet loss rate)
80.3%	73.8%	79.7%	81.2%

Within-subject differences between the conditions were analyzed using one way repeated measures ANOVA. Mauchly's test of sphericity showed that the assumption of sphericity was not violated ($\chi^2(5)=4.207$; $p=.521$). The ANOVA showed that there were differences in the word recall rates between the conditions ($F(3, 57)=3.85$; $p<.05$). In order to compare the conditions without making too many pairwise comparisons three a priori contrasts were selected: Condition 1 (control) vs. Condition 2 (noise), Condition 1 (control) vs. Condition 3 (low packet loss rate) and Condition 1 (control) vs. Condition 4 (high packet loss rate). Results from pairwise t-tests of these a priori contrasts are shown in Table V. The detrimental effect of noise on memory of speech previously shown by Kjellberg et al. [5] was ratified. However, AMR coded speech distorted by packet loss at the rates tested did not show any significant differences from the control condition.

Table V. Within-subject differences in proportions of correctly recalled words, tests of a priori contrasts.

Contrast	Mean Diff.	t(19)	Sig. (2-tailed)	r
Control vs. Noise	6.4%	2.61	.017*	.51
Control vs. Low packet loss rate	0.6%	0.31	.763	.07
Control vs. High packet loss rate	-1.0%	-0.34	.735	.08

*Significant difference. Bonferroni adjustment for multiple comparisons requires $p \leq .017$ for a difference to be significant at 95% confidence level.

4. Discussion

The known detrimental effect of noise on memory of speech [5] was ratified. Sheffield et al. [10] have shown detrimental effects on memory also for low bit rate digital radio speech coding. However, no such effects were seen in this study, even when the transmission was disturbed by high packet loss rates. In Condition 3 (low packet loss rate) the speech intelligibility was comparable to the intelligibility in Condition 2 (noise). In Condition 4 (high packet loss rate) the speech intelligibility was lower than in Condition 2 (noise). Still, no decrease in recall rate was observed when compared to Condition 1 (the undistorted control condition). A possible explanation for this difference between speech distorted by noise and speech distorted by digital speech coding of low quality could be that the digital distortions only affect speech segments, while the noise was evenly distributed over time. The test used in this study allowed subjects to use the pause in between words for rehearsal and for making up strategies for memorization. Noise in these pauses may disturb rehearsal and encoding of perception. Previous studies by Kjellberg et al. [5] and Ljung & Kjellberg [6] on effects of long reverberation time on recall of spoken information have shown that reverberation decreases recall rates only of the earlier parts of long lists of words (50 words), whereas noise affects both early and late parts of such long word lists. This indicates that the effect of noise on recall of the later parts may be the result of distraction caused by the noise during the pauses in between words and not the distortion of the speech signal itself. The effect on the early parts of the lists obtained both under noise and reverberant conditions may be explained by the increase in resource demand caused by noise or long reverberation time, which should leave less time for the transfer to long-term storage and early consolidation in the long-term memory [6]. The use of word lists consisting of 10 words in the present experiment means that only effects shown in the later parts of the long word lists used by Kjellberg et al. [5] and Ljung & Kjellberg [6] can be expected. Since the packet loss distortions have the same character as reverberation, in the sense of only affecting speech passages and not the pauses in between words, the results may be expected. This means that different aspects of disturbances of memory are studied depending on the design of the experiment. From this, the conclusion is that different tests are needed to capture different detrimental effects caused by different kinds of distortions. This experiment could not show that AMR coding or packet loss in AMR coded speech disturb memory of speech, even when high rates of packet loss were introduced. Still, it cannot be concluded that AMR coded speech with packet loss does not disturb memory of speech. Further studies using other kinds of memory tasks should be made to study if these kinds of distortions affect for example transfer to long-term storage.

Memory of speech is an important aspect of speech quality of telecommunication systems. Prediction models for speech quality should therefore include modeling of memory. Since impairment of word identification rates did not correlate with impairment of word recall rates in this experiment, models for quality prediction of speech reproductions cannot be based on identification rates alone. Further, different kinds of memory tasks are required for examination of how different technologies affect speech quality.

5. Conclusions

Memory of speech under three different distorted conditions were compared with an undistorted control condition: speech spectrum weighted noise at 4 dB SNR, AMR coded speech at 12.2 kbit/s with a low packet loss rate (0.5 frames/s) and AMR coded speech at 4.75 kbit/s with a high packet loss rate (0.8 frames/s). The results confirmed the disruptive effect of noise on free recall of words, while no significant impairment was found for the AMR distortions. The noise and the AMR coding with low packet loss rate gave approximately the same impairment of word identification. The AMR coding with high packet loss rate gave a larger impairment of word identification, even though the word recall rate was unaffected. This result suggests that packet

loss in AMR coded telephone systems cause distortions that disrupt recall of words less than noise at levels resulting in the same change of word identification rates. Since impairment of word identification rates did not correlate with impairment of word recall rates models for quality prediction of speech reproductions should not be based on identification rates alone. Further studies are needed to better understand how different kinds of distortions of digitally coded speech affect different memory tasks.

References

[1] R. Ljung, P. Sörqvist, A. Kjellberg, A.M. Green: Poor listening conditions impair memory for intelligible lectures: implications for acoustic classroom standards. Building Acoustics 16 (2009) 257-265.

[2] M.K. Pichora-Fuller: Cognitive aging and auditory information processing. International Journal of Audiology 4 (2003) 26-32.

[3] P.M.A. Rabbitt: Recognition: memory for words correctly heard in noise. Psychonomic Science 6 (1966) 383-384.

[4] P.M.A. Rabbitt: Channel-capacity, intelligibility and immediate memory. Quarterly Journal of Experimental Psychology 20 (1968) 241-248.

[5] A. Kjellberg, R. Ljung, D. Hallman: Recall of words heard in noise. Applied Cognitive Psychology 22 (2008) 1088-1098.

[6] R. Ljung, A. Kjellberg: Long reverberation time decreases recall of spoken information. Building Acoustics 16 (2009) 301-312.

[7] R. Ljung, K. Israelsson, S. Hygge: Speech Intelligibility and Recall of Spoken Material Heard at Different Signal-to-noise Ratios and the Role Played by Working Memory Capacity. Applied Cognitive Psychology 27 (2013) 198-203.

[8] U. Heute: Telephone-Speech Quality. - In: Topics in Speech and Audio Processing in Adverse Environments. E. Hänsler, G. Schmidt (eds.). Springer, Berlin, 2008.

[9] S. Möller, W.-Y. Chan, N. Côté, T.H. Falk, A. Raake, M. Wältermann:. Speech quality estimation: models and trends. IEEE Signal Processing Magazine, 28, 6 (2011) 18-28.

[10] E.G. Sheffield, M.T. Hinch, D.N. Schwab, J.C. Kean: The effects of degraded audio on memory for speech passages. IEEE Transactions on Broadcasting 55 (2009) 569-576.

[11] B. Hagerman: Sentences for testing speech intelligibility in noise. Scandinavian Audiology 11 (1982) 79-87.

[12] J. Sjöberg, M. Westerlund, A. Lakaniemi, Q. Xie: RTP payload format and file storage format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) audio codecs, IETF RFC4867, 2007.