Evaluation of Speech Synthesis Systems using the Speech Reception Threshold Methodology

David A. van Leeuwen and Johan van Balken
TNO Human Factors
Postbus 23
3769 ZG Soesterberg
The Netherlands

{vanLeeuwen,vanBalken}@tm.tno.nl

ABSTRACT

The intelligibility of speech sysnthesis systems that are available nowadays is usually high enough to enable comparisons between different synthesis systems based on the speech quality. However, in some situations, like a civil aircraft cockpit, the acoustic environment may be such that intelligibility is a discriminating factor between systems. In this paper we propose a methodology for comparing speech synthesis systems based on the Speech Reception Threshold (SRT). With this method the signal-to-noise ratio is found at which 50% intelligibility of redundant sentences is reached. A system with a lower SRT value is said to be more robust against masking noise. We have compared 5 commercial speech synthesis systems (4 male voices, 5 female voices) in an SRT experiment using a masking noise that was spectrally equivalent to cockpit noise. SRT values range from -4.1dB to 1.1dB. An ANOVA revealed that two of the nine systems had a significantly lower SRT than the rest. There was also an effect of the test subject, which is remarkable because the SRT has usually small variability over listeners.

INTRODUCTION

The EU 6th framework programme project SAFESOUND involves research to implement modern speech and audio technologies in the cockpit of a civil airline aircraft, in order to enhance the safety of flights. One of the subsystems implements Direct Voice Output (DVO) in order to communicate system messages to the pilots. For the DVO system a commercial-off-the-shelf (COTS) speech synthesis system is envisaged. The main selection criterium for the system is maximum intelligibility of the messages in the noisy civil aircraft cockpit.

The quality of commercially available speech synthesis systems is considered to be high enough that for most applications the *intelligibility* of the systems is not an issue, but rather other performance measures such as speech quality or listening comfort. However, for applications that should increase safety in a noisy cockpit environment, we argue that intelligibility *is* of highest concern in the choice of a synthesis system, since the mis-interpretation of a spoken system message can have negative consequences to the safety of the aircraft. Several methodologies for assessing the intelligibility can be found in literature, e.g., diagnostic rhyme test (DRT, see Voiers, 1977), or the Diphone Test (Pols et al., 1987; van Bezooijen and Pols, 1987). For a good overview of these methodologies, see van Bezooijen and van Heuven (1997). These methodologies were developed at a time when the technology of speech synthesis systems was not as advanced as it is nowadays. Then, typical systems were formant-based or diphone-based, while modern systems are based on unit selection. In these methodologies ambient noise is not the key factor for determining the intelligibility, but rather the specific technology that is used. We therefore investigated new methodologies and came up with extending the use of the Speech Reception Threshold (SRT) to speech synthesis systems.

The SRT methodology uses short redundant sentences as speech stimulus material that subjective listeners are exposed to. The sentences are linguistically meaningful, and consist of 8-9 syllables. This type of

van Leeuwen, D.A.; van Balken, J. (2005) Evaluation of Speech Synthesis Systems using the Speech Reception Threshold Methodology. In *New Directions for Improving Audio Effectiveness* (pp. 9-1 – 9-6). Meeting Proceedings RTO-MP-HFM-123, Paper 9. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.aps.

sentence is better suited to the modern unit selection based synthesis system than more diagnostic linguistic material, such as Semantically Unpredictable Sentences (SUS) or nonsense consonant-vowel-consonant (CVC) words or variants thereof. The SRT determines the signal to noise ratio (SNR) at which 50% of the test sentences are perceived correctly by test subjects. The measure can be determined accurately using a relatively low number of listeners, because the variance within (normal hearing native) listeners is small. Because the psychometric curve that describes the relation between fraction of correctly recognized sentences and SNR is rather steep, the SRT value is a good measure for the minimum SNR needed for proper intelligibility of messages. Therefore, the SRT value for a speech synthesis system in a noisy environment is directly related to the sound level at which synthesized messages need to be played in order to be perceived correctly. In a civil aircraft cockpit, with many different sound sources and a non-negligible background noise level, it is desirable to keep the sound level of these synthesized messages low, implying the need for a low SRT value.

EXPERIMENTAL SETUP

In an SRT experiment (Plomp and Mimpen, 1979) a test subject listens to a list of 13 recorded sentences one by one. The speech signal of each sentence is masked by noise, and the level of the noise (or signal to noise ratio) varies within the list. The subject is requested to orally repeat the sentence. The first sentence in a list is presented at a very low SNR, and repeated with increasing SNR until it is repeated correctly. In the next steps a new sentence is presented irrespective of the listeners response, until all 13 sentences have been played. If the sentence is repeated correctly, the noise level is increased, and if one or more words is not correct the noise level is decreased. The SRT-level is defined as the average signal-to-noise ratio over the last 10 sentences of the list. If the standard deviation of the last 10 SNR values exceeds 3dB, the SRT value is considered invalid and the data point is not used in subsequent analysis. The speech level is determined using the `speech level meter' (SLM) algorithm (van Velden, 1991), which accounts the root-mean-square level of the upper 14dB of the speech level histogram.

We used American English SRT sentences applied previously in non-native speech research (van Wijngaarden, 2001). The sentences were transformed to speech by 5 different COTS systems, with both male and female voices resulting in 9 voice conditions (one supplier had only a female voice available). The systems voices were also of American English accent. The systems were used `out of the box', and no adaptation of the sentences to particular systems was applied. An exception was made for the word `aeroplane' used in one of the sentences, which several synthesis systems could not pronounce properly. This word was changed to the less British `airplane.'

A panel of 10 Dutch subjects participated in the SRT test in a 10×10 Latin square design. In total we used 10 different lists of 13 sentences, so that for each synthesized voice 130 sentences were produced. Because there was one voice condition less than the number of subjects or sentence lists, the Latin square design skipped one list of sentences for each subject. The masking noise used for the SRT was spectrally equivalent to noise in a Airbus A320 cockpit recorded at the location of the flying pilot during climbing of the aircraft. This stage of the flight has a higher noise level in the cockpit than during other stages, so it was chosen to represent the most critical condition with respect to intelligibility.

RESULTS

In Figure 1 the distributions of SRT over listeners are depicted for the various voices (systems). The measured average SRT values for the nine systems lie in the range -4.2 dB to $+1.1$ dB. These results are in the same range as normal SRT values of -3 dB for undistorted wideband speech of human

speakers with native listeners, and to SRTs around +1 dB for Dutch subjects listening to native English speakers (van Wijngaarden, 2001). The better systems appear to score as well as human speakers in the non-native condition for the SRT, but we must remark here that in those SRT experiments the masking noise is shaped according to the long-term average spectrum of the speaker, thus providing optimal masking. The spectrum of the noise used in this experiment has most energy in the lower frequencies, rolling off with an approximate slope of 3dB/octave. Both speech signal and noise spectrum have been A-weighted for determining the SNR.

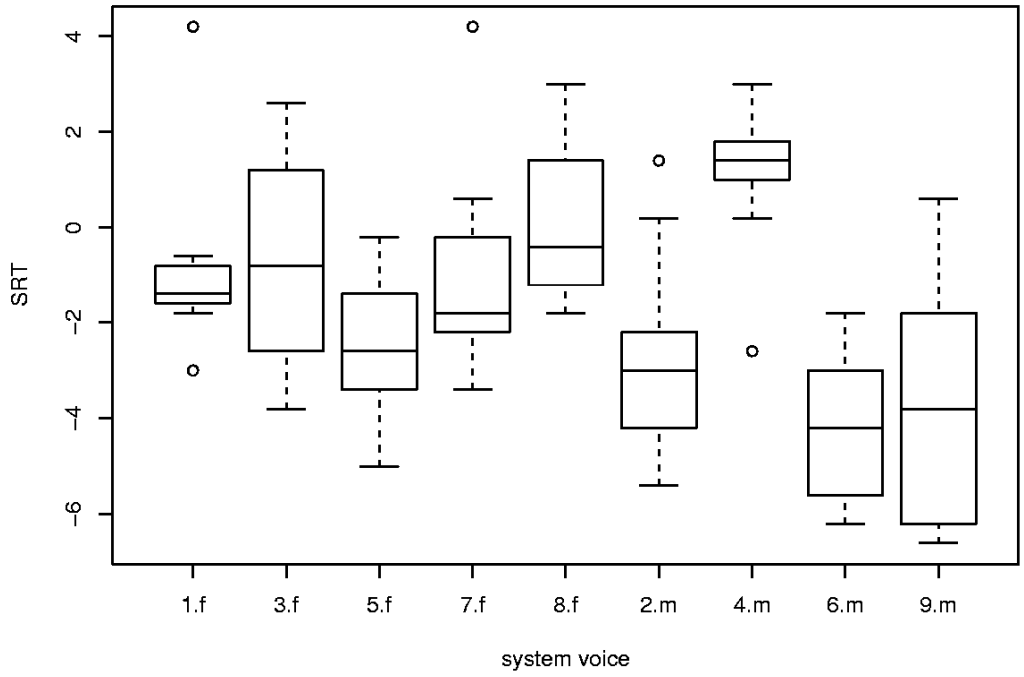

Figure 1: Boxplots of the measured SRT values averaged over synthesis system. A box indicates the first to third quartile of the distribution of SRT value among subjects. The horizontal line represents the median. The `whiskers' indicate the range of remaining data points, except for extremes which are plotted individually. The gender of the system's voice is indicated in the system label as `.m' or `.f'.

Of the 90 SRT data points (9 systems × 10 subjects), 78 were valid. The results of the experiments have been analyzed in an analysis of variance (ANOVA). The results for the main effects (`system' and `subject') have been tabulated in Table 1.

Table 1: SRT Anova results of main effects

Factor	F	p
Speech synthesis system	12.6	2×10^{-10}
Subject	8.4	5×10^{-8}

There are clear main effects for the factors `system' and `subject'. A pairwise t-test shows that only systems 4, 6 and 9 have SRT values significantly different ($p < 0.05$) from most other systems. It is interesting that the all these systems use a male voice. Despite the wide range of SRT values for systems with male voice, a t-test between male voice and female voice data reveals a significant difference in SRT means of 1.3dB ($p = 0.024$).

As can seen from Figure 1 the sizes of the boxes, indicating the 25th and 75th percentile, are generally large. This suggests a large spread in SRT values over the different subjects. In order to test if this is systematically due to the subjects, the means of SRT values over systems for the different test subjects are plotted in Figure 2. The means over system range from -3.2 dB to 0dB with an outlier at 1.6 dB for subject 10. This is more variability than expected, because the SRT method normally yields results with small standard deviations within subjects (Plomp and Mimpen, 1979). The observed variance within listeners, $\sigma = 2$dB might be due to the inherent non-native character of the experiment. Van Wijngaarden et al. (2002) report a $\sigma = 1$dB. In this case, subject 10 might have been less familiar with listening to the English language than the others. Although the subjects were asked if they felt they would be able to understand the English well enough, their English proficiency was not explicitly measured.

CONCLUSION

The experiment shows that the methodology of using the SRT with speech synthesis systems can be useful in an applied scenario such as an aircraft cockpit, where ambient noise is an important operational factor. An noteworthy difference with traditional SRT measurements is that the masking noise is not spectrally shaped like the long-term spectrum of speech, but according to the actual noise at the application domain. This leads to SRT values which are not directly comparable to other values reported in literature, but we argue they *are* meaningful to the specific application.

The power of the current experiment seems to be a little too low to make a well-funded choice for a particular speech synthesis system. With this first experiment we can estimate the power needed for subsequent experiments. We argue that more investigation of the non-native factor is necessary given the observed high listener variability and the international use of English-centered aviation equipment.

It is interesting that synthesized male voices have lower SRT values (i.e, higher intelligibility in an noisy environment) than female voices. There are several ways of reasoning about this difference. The traditional explanation for this is that with a higher fundamental frequency the female voices lead to less filled formants, and therefore less discriminable vowels, than their male counterparts. For the reduntant sentences used in the SRT experiments the intelligibility in noise might be based mainly on the recognition of these vowels with only timing information from the consonants, which themselves are masked by the noise. Another cause may be sought in the relative position of the male and female speech spectra with respect to the masking noise, and the A-weighting of the different octave bands for applying a particular SNR. For example, the 125Hz octave band is known to contibute to the intelligibility, but does hardly count in the A-weighting. The difference between female and male speech in this band is significant.

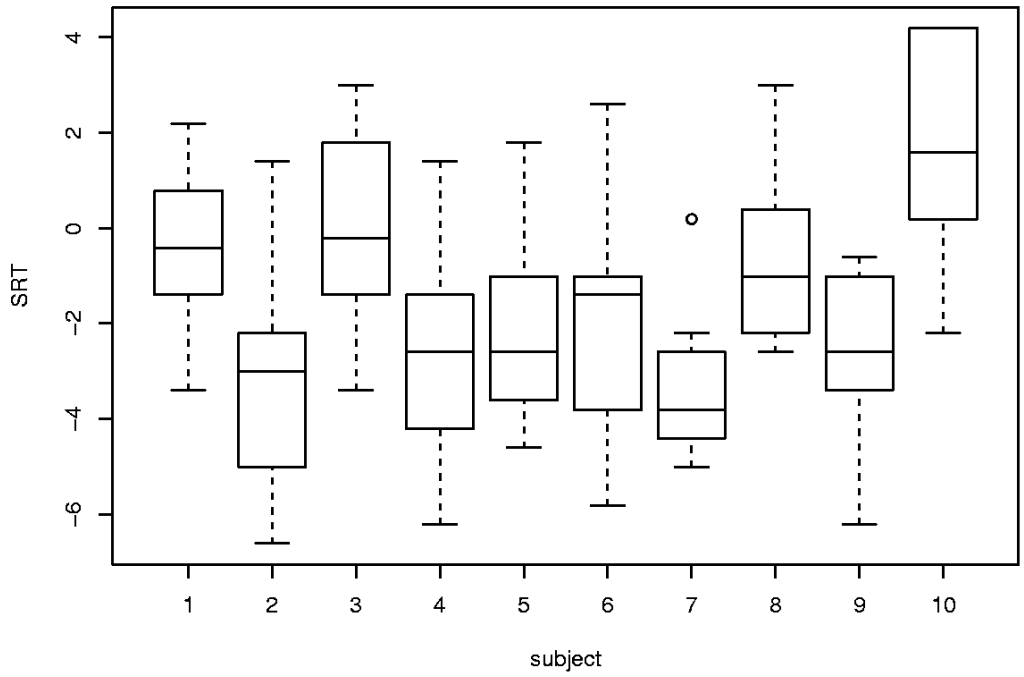

Figure 2: Boxplots of the SRT values averaged over different listeners.

Given the large spread in measured SRT values among systems, it appears that the intelligibility in noise is particularly dependent on the voice of the system. For instance, systems 8 and 9 are the female/male counterparts of the same manufacturer, and therefore share the same synthesis technology. Yet they show a large difference in SRT value, about 3.6dB.

Perhaps, referring to the well-known Turing test, we might claim that the quality of speech synthesis systems has really matured when they consistently score the same SRT values that humans do...

BIBLIOGRAPHY

[1] R. Plomp and A. M. Mimpen. Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 8:-52, 1979.

[2] L. C. W. Pols, J.-P. Lefèvre, G. Boxelaar, and N. van Son. Word intelligibility of a rule synthesis system for french. In *Proc. Eurospeech*, volume 1, pages 179-182, 1987.

[3] R. van Bezooijen and L. C. W. Pols. Evaluation of two synthesis-by-rule systems for dutch. In *Proc. Eurospeech*, volume 1, pages 183-186, 1987.

[4] R. van Bezooijen and V. van Heuven. *Handbook of Standards and Resources for Spoken Language Systems*, chapter Assessment of synthesis systems, pages 481-563. Mouton de Gruyter, 1997.

[5] J. G. van Velden. Speech Level Meter, vesion 2, SAM_SLM. Technical report, TNO Human Factors, Soesterberg, 1991. SAM-TNO-043.

[6] S. J. van Wijngaarden. Intelligibility of native and non-native Dutch speech. *Speech communication*, 35:-113, 2001.

[7] S. J. van Wijngaarden, H. J. M. Steeneken, and T. Houtgast. Quantifying the intelligibilty of speech using non-native listeners. *J. Acoust. Soc. Am.*, 111:-1916, 2002.

[8] W. D. Voiers. Diagnostic evaluation of speech intelligibility. In M. E. Hawley, editor, *Benchmark papers in Acoustics*, number 2, pages 374-384. Dowden, Hutchinson, and Ross, Stroudburg, 1977.