Automating speech reception threshold measurements using automatic speech recognition

Hanne Deprez[1]*, Emre Yılmaz*[1]*, Stefan Lievens*[2]*, Hugo Van hamme*[1]

[1] Dept. of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[2] Cochlear Technology Center Belgium, Schaliënhoevedreef 20i, Mechelen, Belgium

hanne.deprez@student.kuleuven.be, emre.yilmaz@esat.kuleuven.be,
slievens@cochlear.com, hugo.vanhamme@esat.kuleuven.be

Abstract

The speech reception threshold (SRT) is the noise level at which the speech recognition rate of a test person is 50%. SRT measurement is relevant for patient screening, psychoacoustic research and algorithm development in hearing aids and cochlear implants. In this paper, we report on our efforts to automate SRT measurement using an automatic speech recognizer. During a test, sentences are presented to the test subject at different SNR levels. The person under test repeats the sentence and the keywords it contains are scored by an audiologist. If all keywords are repeated correctly, the sentence is evaluated as correct. The SNR level of each sentence is adjusted based on the previous sentence's evaluation. Aiming for an objective and repeatable measurement, the audiologist's assessment is replaced by an automatic speech recognizer's evaluation. For this purpose, we investigate different finite state transducer structures to model the expected sentences as well as the impact of several speaker adaptation schemes on the keyword detection accuracy. A baseline recognizer using general acoustic models achieves a performance of 88.8% keyword detection rate. Speaker adapted acoustic models improve the performance yielding a keyword detection accuracy of up to 90.7%. Finally, the impact of recognition errors on the estimated SRT value is simulated showing a minimal impact on the SRT measurement process. Based on this analysis, it can be concluded that the proposed automatic evaluation scheme is a viable tool for speech reception threshold measurements.

Index Terms: keyword detection, speaker adaptation, cochlear implant, speech test, speech reception threshold

1. Introduction

Speech reception threshold (SRT) measurements have been used in a clinical setting for evaluating a person's hearing capabilities and to diagnose hearing loss. The obtained SRT value is a subjective measure for quantifying the hearing ability of patients with cochlear implants (CI) in order to adjust the CI parameters and analyze the impact of new developments in CI devices on the patient's hearing abilities [1, 2, 3]. Moreover, these measurements provide useful data for psychoacoustic research, e.g. to investigate how cognitive load influences speech recognition of individuals.

Several Dutch speech tests for determining a patient's speech recognition threshold have been proposed, e.g. NVA-tests [4] and LIST-tests [5]. During these tests, words or sentences which are embedded in different levels of noise are presented to patients and they are asked to repeat what they hear. The responses are evaluated by an audiologist who decides if patients properly repeat the presented word or sentence. LIST-tests consist of ten sentences that are presented to a patient at a certain noise level. For each sentence, two to five content words (called keywords henceforth) are defined. Each keyword in the patient's response is evaluated by the audiologist and if all keywords were reproduced correctly (incorrectly), the noise level in which the following sentence is embedded is increased (decreased) by 2 dB resulting in a more (less) challenging recognition task. After ten sentences, the SRT value is obtained by averaging the SNR levels at which the last six sentences are presented. This speech reception threshold corresponds to the point where 50% of the keywords are understood correctly by the patient.

At the outset of this study, the SRT test procedure was identified as one that was particularly apt for automation since it seems feasible to set up an automatic speech evaluation method that makes significantly fewer errors than the human under test, who operates around a 50% rate. Hence, errors introduced by the speech recognizer are expected not to affect the test outcome significantly. An automated test provides the additional benefit of an objective and repeatable measurement compared to an audiologist whose evaluation may be biased. Furthermore, automating this procedure saves a great amount of time in which audiologists could focus more on their core tasks: providing a better assistance to CI patients.

Automation of these tests was investigated in [6] by letting the patients type what they have heard while accounting for spelling errors. A rehabilitation tool for CI users using automatic speech recognition (ASR) is described in [7]. CI patients are encouraged to repeat spoken sentences upon which correctness feedback is provided using ASR. The proposed system for SRT tests is similar in recognition task, but differs in the language model constraints since the main task is to detect the keywords rather than recognition of the complete utterance. It also differs from traditional keyword spotting (KWS) [8, 9, 10] because the knowledge of the embedding sentence can be exploited while KWS is mainly used for unconstrained and spontaneous speech. As the expected utterances are known in the scope of this paper, the use of deterministic language models is feasible. The design procedure of these deterministic language models is presented further in this paper.

We have further investigated the impact of several speaker adaptation techniques on the keyword detection accuracy. In this scenario, the data of an earlier SRT measurement session with the same patient is reused to adapt his/her acoustic models. Several adaptation methods such as MLLR [11] and constrained and unconstrained linear mean and covariance transforms [12] are applied to the speaker independent acoustic models and the

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 35–40,
Grenoble, France, 21–22 August, 2013.

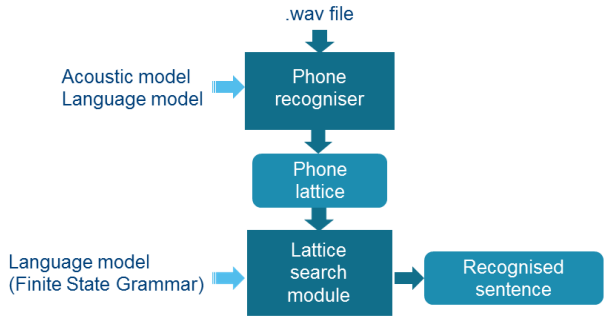

Figure 1: *Two layered speech recognition architecture.*

performances of the adapted models are compared.

The rest of the paper is organized as follows. Section 2 introduces the speech recognizer's architecture and discusses the design of deterministic language models and the speaker adaptation techniques that are applied in the experiments. The experimental setup is described in Section 3 and the results are presented in Section 4. Finally, the conclusions are discussed in Section 5.

2. Automatic Speech Evaluation Scheme

The proposed evaluation scheme uses an automatic speech recognizer that replaces the audiologist during the SRT measurements. The overview of the ASR that is used for this purpose is given in Section 2.1. As the expected utterances are known, deterministic language models with different structures are designed and used during recognition. Section 2.2 details the design procedure. Finally, several speaker adaptation techniques are applied to investigate the impact on the recognition accuracy which is the topic of Section 2.3.

2.1. ASR overview

A two layered HMM-based recognition system as illustrated in Figure 1 is used for obtaining the word-level recognition output. In the first layer, a phone recognizer generates a phone lattice using task-independent acoustic and language models. These models can be general models that are trained on the data of the target language. In the second layer, task-dependent information is provided in the form of a finite state transducer (FST) describing lexical and grammatical knowledge. The FST is composed of two levels, namely the word and garbage FSTs modeling the phone level information and the sentence FST containing multiple word and garbage FSTs to model the expected utterances. This structure comes with increased modularity as the generic phone recognizer can be used for any recognition task provided that the task-specific information is contained in the second stage [13]. Using the task-dependent information incorporated in the FSTs, the phone lattice obtained in the previous step is decoded into a word level recognition result which can further be processed to obtain the keywords that have been uttered.

2.2. Language model design

The basic FST structure models the expected sentence by allowing the correct utterances of the words in the order they appear in the prompt. Incorrect or irrelevant utterances are modeled by the garbage FST. However, due to the nature of SRT measurement tests, it is a requirement to have higher flexibility in the sentence FST as the patients can repeat the presented words in arbitrary order or they may skip some of the presented words. All FSTs consist of a number of nodes and arcs depending on the number of phones and words in the expected sentence. The start and end nodes are marked with <s> and </s> respectively. All other nodes are labeled with the keywords: visiting a state indicates the associated keyword was detected. State transitions occur upon a match between a word or phrase model (the edge's earmark) and a partial path in the phone lattice output by the first layer. Non-keywords (henceforth *filler words*), silence (marked with #) and garbage (marked with GBG) cause a self-transition. Garbage models any unanticipated speech allowing any phone sequence. To keep it from being preferred over other edges, it is penalized with a *garbage model cost* that is incurred once upon entry. Based on this principle, three different FSTs are designed modeling the expected patient's response, each of which handles the filler words differently.

In the first model, named the KWandFILLER model, each filler word is accepted as an input with an arc linked to the node of the preceding keyword. This model is illustrated with an example for the Dutch sentence "MAMA vertelt ons elke AVOND een kort VERHAAL" (MOM reads us a short STORY every NIGHT) in Figure 2, where keywords are written in uppercase characters.

In the KWandLONGFILLER model, only filler words of sufficient length are added to the model in order to limit the number of falsely detected filler words. This model is expected to reduce the false alarms due to short filler words.

The third design, the KWandFILLERSEQ model, contains a single arc that is associated with all filler words that appear between two keywords. In this model, the canonical order of the filler words is taken into account. This could have the advantage that the filler words are recognized in the correct order and should prevent (especially short) fillers from erroneously modeling keywords.

2.3. Speaker adaptation techniques

Speaker adaptation is implemented by linearly transforming the means and possibly also the covariances of the Gaussians of a speaker independent (SI) acoustic model. This transform is obtained by maximizing the likelihood of a selection of adaptation data as described in [11] and [12].

Three different adaptation techniques, namely a linear mean transform (MLLR), constrained mean and covariance transform (CMLLR) and unconstrained mean and covariance transform (UMLLR), are investigated. For MLLR, the means (μ) of the Gaussians of the SI acoustic models are linearly transformed with a transformation matrix W: $\hat{\mu} = W\xi$ with $\xi = [1\ \mu]$. For UMLLR, the transformation matrix W of the means (μ) and the transformation matrix H of the covariances (Σ) are separate: $\hat{\mu} = W\xi$ and $\hat{\Sigma} = H\Sigma H^T$. In the case of CMLLR, the transformation A' applied to the variances (Σ) must correspond to the transformation A' applied to the means (μ): $\hat{\mu} = A'\mu - b'$ and $\hat{\Sigma} = A'\Sigma A'^T$. These transforms are obtained by maximizing the likelihood of the adaptation data, details of which are given in [11] and [12].

In each of these adaptation schemes, the states that are present in the adaptation data should be provided. This information is captured in a state segmentation which is generated from a transcription of the utterance. This transcription is acquired by manual annotation of the data. To avoid this manual

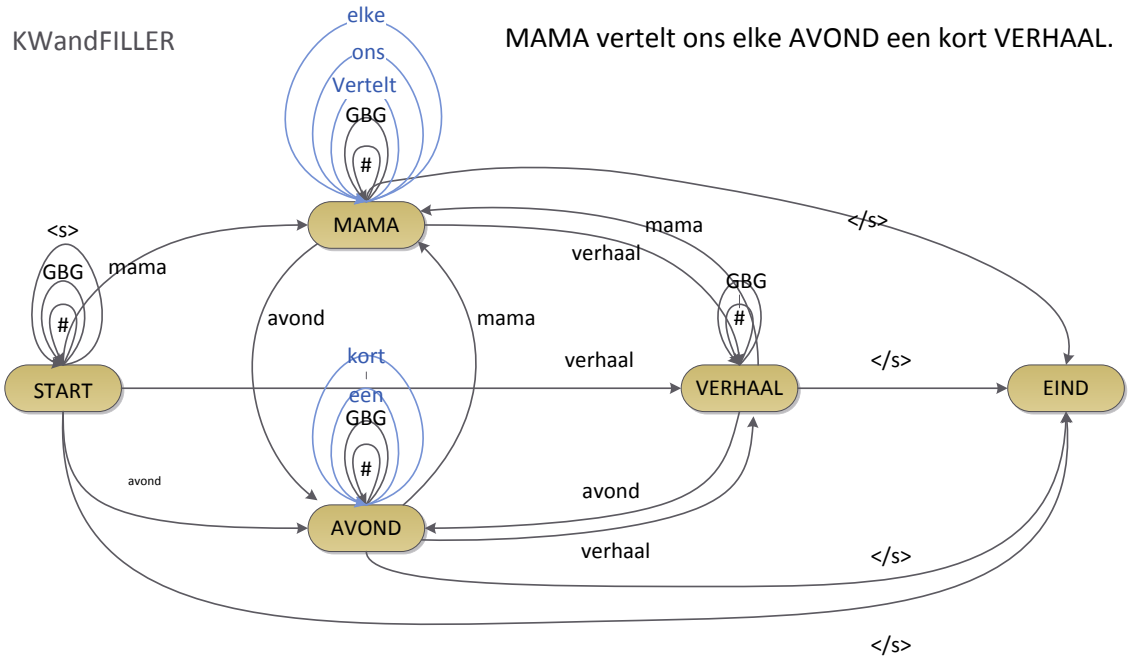

Figure 2: *Example of the KWandFILLER FST model.*

intervention, unsupervised adaptation is also considered, where only sentences that were assessed as correct by the system are retained as adaptation data.

3. Experimental setup

3.1. Speech data and baseline recognizer

The performance of the baseline system with the presented FST models and of the system with the adapted acoustic models was evaluated on recordings that contain the patient's responses to LIST-tests performed by normal hearing persons. Utterances from 17 speakers two of which are non-native Dutch speakers are captured in a recording cabin used for SRT measurements. In total, 79 lists are evaluated resulting in 4.64 lists per person on average. For the speakers with enough recorded lists, speaker adaptation was applied and performance of the speaker adapted system is evaluated using cross validation to obtain statistically significant results.

The acoustic models were trained based on the CoGen database ([14]) which contains 7 hours of read speech. The speaker independent acoustic models are semicontinuous HMMs with tied Gaussians consisting of 576 states and 10635 Gaussians. The task-independent language model consists of a trigram phoneme sequence model derived from a Dutch database with correctly read sentences [15]. The preprocessing is based on Mel-spectrum analysis and includes cepstral mean subtraction and discriminant analysis (MIDA) [15] [16].

3.2. Evaluation metrics

When evaluating the quality of the automated CI test, there are two important errors to consider: not detecting correct sentences on the one hand and classifying a sentence that is incorrect as correct on the other hand. Two performance criteria have been defined: keyword detection rate (KDR) quantifying the former and false alarm rate (FAR) quantifying the latter. Both of these metrics are defined at the sentence level, since the SNR is adapted based on the evaluation of an *entire* sentence. A sentence is correct if all keywords are repeated correctly by the patient and incorrect if the patient missed at least one keyword.

$$\text{KDR} = \frac{\#\text{ of correctly detected sentences}}{\#\text{ of correct sentences}} \quad (1)$$

$$\text{FAR} = \frac{\#\text{ of sentences incorrectly classified as correct}}{\#\text{ of incorrect sentences}} \quad (2)$$

4. Results and discussion

4.1. Baseline system

The FSG models presented above are evaluated according to their performance by means of a KDR-FAR plot in Figure 3. There are three different operating points obtained by manipulating the phone lattice density. The equal error rate points are marked with $\Diamond$. The KWandLONGFILLER model provides the worst performance, whereas the other two models perform similarly. The reason for the bad performance of the KWandLONGFILLER model is that it has to use the garbage model to model the short filler words. The performance of the model is thus very dependent on the choice of the garbage model cost. If the garbage model cost is very high, keywords might be detected at the instants where short filler words are uttered. On the other hand, if the garbage model cost is too low, the garbage model is often used to explain the utterance resulting in an increased number of keyword deletions. The performance of the KWandFILLER and KWandFILLERSEQ model are comparable. The KWandFILLER model is the most flexible of the two allowing patients not to say filler words or repeat them in any order, though such deviations do not occur often in our data. Since it is expected that the KWandFILLER model would per-

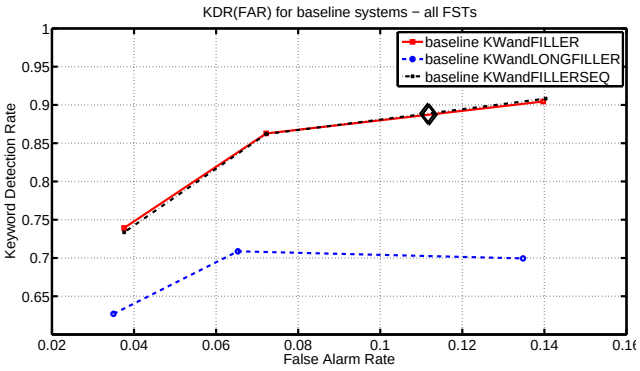

Figure 3: *Comparison of different FST models for the baseline system.*

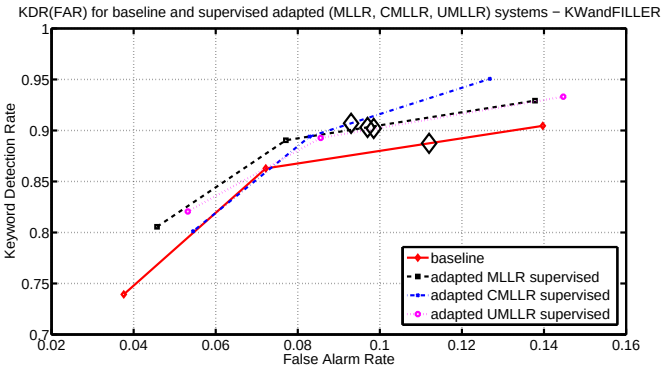

Figure 4: *Comparison of the adapted system performance (MLLR, CMLLR and UMLLR) with baseline system using the KWandFILLER model.*

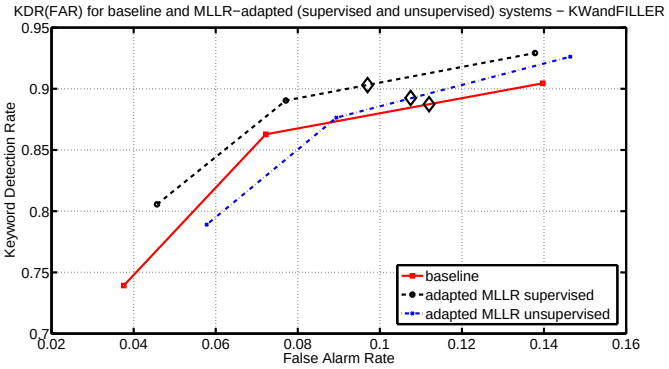

Figure 5: *Comparison of the MLLR-adapted system performance (supervised and unsupervised) with baseline system using the KWandFILLER model.*

form better in case a patient would deviate from the canonical word order, the KWandFILLER model is the best choice for practical applications. The equal error rate point is at a FAR of 11.2% and a KDR of 88.8% as indicated in Figure 3.

4.2. Speaker adapted system

The three adaptation techniques described above are implemented and the obtained KDR-FAR curves are illustrated in Figure 4. The adapted systems perform better than the baseline at most of the operating points. The equal error rate point is obtained at a false alarm rate of 9.7% for MLLR, 9.85% for UMLLR and 9.3% for CMLLR as indicated in the figure.

These adapted models are obtained using the manually annotated adaptation data from two LIST-tests (20 sentences). The adapted models for a certain speaker were tested on the other recorded lists for that speaker. To obtain enough statistical relevance, cross-validation is applied.

In the case of unsupervised adaptation, only sentences which were evaluated as correct by the baseline recognizer are included as adaptation data. When considering two lists per person, only a limited number of adaptation sentences could be included. It was not possible however to consider more lists, because of the limited number of recorded lists per speaker. Here, the expected utterance is used as the transcription. In Figure 5 the KDR-FAR curves for baseline, supervised and unsupervised adapted systems are plotted. The adaptation technique that was applied here is MLLR. The unsupervised adapted system performs worse than the baseline at some operating points. This is because not enough adaptation data could be included, due to the limited number of recordings per person. The equal error rate point for the unsupervised adapted system is obtained at a false alarm rate of 10.75%, compared to the 9.7% FAR for the supervised adapted system.

4.3. Theoretical impact of the recognition error on the measured SRT-value

A LIST test consists of ten sentences, the first of which is presented at a very low SNR. This sentence is repeated until it is evaluated as correct. Then, we advance to the next sentence adapting the SNR at which the sentence is presented according to the evaluation of the previous sentence. In the end, the mean of the SNR at which the last six sentences were presented is taken as the measured SRT-value.

Since the recognizer makes errors by not detecting correct sentences and falsely evaluating incorrect sentences as correct, the measured SRT using the automatic procedure will deviate from the manually obtained value. The effect of the recognizer error on the final SRT is modeled using performance intensity functions. These performance intensity functions model the patient's score as a function of the SNR at which the sentence is presented. An example of a performance intensity curve is given in Figure 6. Based on the input SNR, the probability of a patient understanding the sentence correctly is determined. A binomial variable with this probability is drawn indicating the patient's evaluation of the sentence. A recognition error is introduced by the speech recognizer which may flip this evaluation adjusting the SNR in the wrong way. Based on the recognizer's evaluation, the next SNR is calculated. By simulating a large number of lists, we obtain the distribution of the measured SRT-value with and without a recognizer error. Without introducing the recognizer error, the mean measured SRT over 300 lists is found to be -7.8 dB with a standard deviation of 1.2 dB. With a recognizer error of 10 %, the mean measured SRT becomes -8.0 dB with a standard deviation of 1.8 dB. The evolution of the mean and standard deviation of the measured SRT in function of the ASR's error rate are presented in Figure 7 and 8 respectively. It can be seen that the mean measured SRT value deviates further from the initial value of -7.8 dB for normal hearing persons as the recognizer error increases. The standard deviation on the

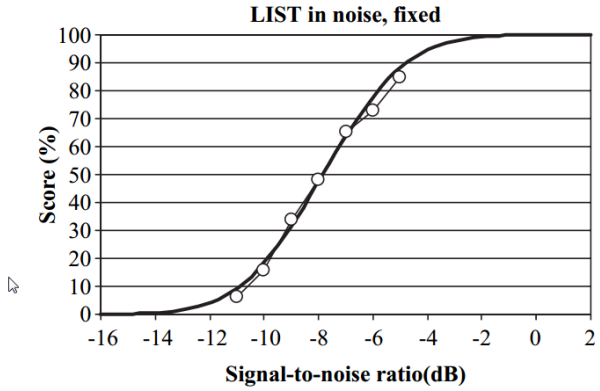

Figure 6: *Performance intensity curve for a LIST sentence presented at a certain SNR. (Taken from [5]).*

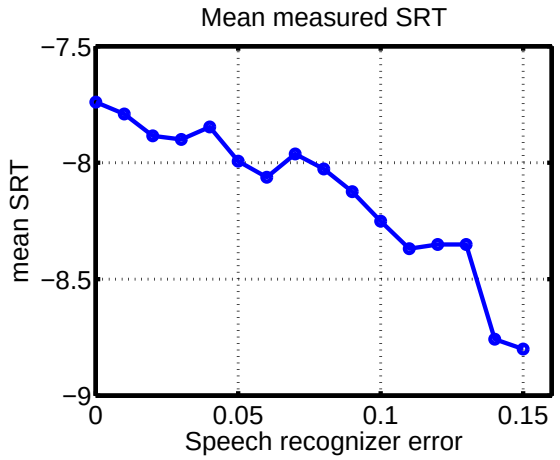

Figure 7: *The mean of the measured SRT as a function of the speech recognition error.*

measured SRT also increases with an increase in the recognizer error.

When new CI techniques are assessed, a comparative measurement before and after activation of the new component is performed. In this case, the bias on the measurement observed when comparing the manual and the automatic test results is of minor importance. It is important however that measurements can be conducted with significant accuracy. If desired, the standard deviation on the measured SRT can be reduced using more sentences per LIST-test. Using 20 instead of 10 sentences per LIST, reduces the standard deviation on the measured SRT to 1.13 dB, for a recognizer error of 10%.

Another use of LIST-tests is to assess the hearing of patients based on their SRT score. In this task, an absolute SRT value is obtained and hence a bias might lead to inaccurate estimations. However, when assessing whether a person has normal hearing or needs some treatment, the differences in SRT scores are so large that this bias will not lead to a different evaluation.

5. Conclusions

A Dutch CI speech reception threshold test (LIST) has been automated using automatic speech recognition. The LIST consists

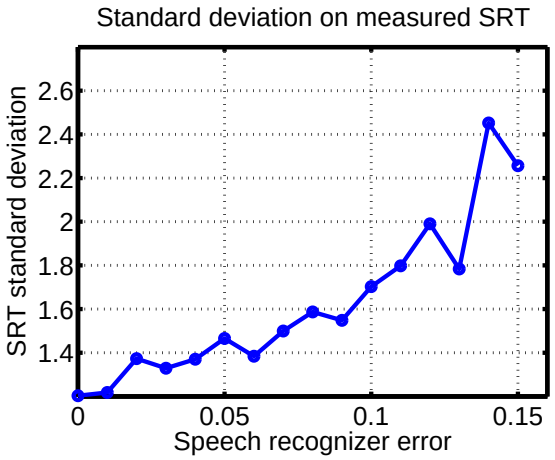

Figure 8: *The standard deviation on the measured SRT as a function of the speech recognition error.*

of ten sentences played at different SNR levels depending on the evaluation of the previous sentence. The speech reception threshold is estimated as the mean of the last six SNR levels.

A speaker independent speech recognizer can work at an operating point with a false alarm rate of 11.2% and keyword detection rate of 88.8% which are both defined at the sentence level. Speaker adaptation improves the results to 9.3% false alarm rate and 90.7% keyword detection rate. The results are obtained at the equilibrium point on the keyword detection rate-false alarm rate curve which reduces the impact of recognition errors on the measured SRT value.

Furthermore, a simulation of the impact of recognizer error on the SRT estimate is provided. In comparison to a manually performed test, there is a bias of 0.2 dB on the SRT measured with the automatic procedure. The standard deviation also increases from 1.2 dB to 1.8 dB. We conclude that these results are sufficiently small for using the automated test in practice.

6. Acknowledgements

The authors would like to thank the participants of the recording sessions of the LIST-tests.

7. References

[1] P. C. Loizou, O. Poroy, and M. Dorman, "The effect of parametric variations of cochlear implant processors on speech understanding," *The Journal of the Acoustical Society of America*, vol. 108, p. 790, 2000.

[2] J. Müller, F. Schon, and J. Helms, "Speech understanding in quiet and noise in bilateral users of the MED-EL COMBI 40/40+ cochlear implant system," *Ear and Hearing*, vol. 23, no. 3, pp. 198–206, 2002.

[3] M. F. Dorman, P. C. Loizou, and D. Rainey, "Simulating the effect of cochlear-implant electrode insertion depth on speech understanding," *The Journal of the Acoustical Society of America*, vol. 102, p. 2993, 1997.

[4] J. Wouters, W. Damman, and A. J. Bosman, "Vlaamse opname van woordenlijsten voor spraakaudiometrie," *Logopedie: informatiemedium van de Vlaamse vereniging voor logopedisten*, vol. 7, no. 6, pp. 28–34, 1994.

[5] A. Van Wieringen and J. Wouters, "LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands," *International journal of audiology*, vol. 47, no. 6, pp. 348–355, 2008.

[6] T. Francart, M. Moonen, and J. Wouters, "Automatic testing of speech recognition," *International Journal of Audiology*, vol. 48, no. 2, pp. 80–90, 2009.

[7] W. Nogueira, F. Vanpoucke, P. Dykmans, L. De Raeve, H. Van Hamme, and J. Roelens, "Speech recognition technology in CI rehabilitation," *Cochlear Implants International*, vol. 11, no. Supplement 1, pp. 449–453, 2010.

[8] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 129–132.

[9] H. Bourlard, B. D'hoore, and J.-M. Boite, "Optimizing recognition and rejection performance in wordspotting systems," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1. IEEE, 1994, pp. I–373.

[10] R. C. Rose, "Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition," *Computer Speech & Language*, vol. 9, no. 4, pp. 309–333, 1995.

[11] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[12] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, 1998.

[13] J. Duchateau, M. Wigham, K. Demuynck, and H. Van hamme, "A flexible recognizer architecture in a reading tutor for children," in *Proc. of the ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006, pp. 330–331.

[14] K. Demuynck, D. Van Compernolle, C. Van Hove, and J.-P. Martens, "CoGen een corpus gesproken Nederlands voor spraaktechnologisch onderzoek - eindverslag," *Tech. Rep. K.U. Leuven - ESAT & Universiteit Gent*, 1997.

[15] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst *et al.*, "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Communication*, vol. 51, no. 10, pp. 985–994, 2009.

[16] K. Demuynck, J. Duchateau, and D. Van Compernolle, "Optimal feature sub-space selection based on discriminant analysis," in *Proc. Eurospeech*, vol. 3, 1999, pp. 1311–1314.