Human and machine recognition as a function of SNR

Bernt Andrassy* **and Harald Hoege**

Siemens AG, Corporate Technology, IC 5 Munich, Germany
{firstname.lastname}@siemens.com

Abstract

In-car automatic speech recognition (ASR) is usually evaluated by determining one single word error rate (WER) for an in-car task. This measure does not allow to look at the recogniser behaviour for different levels of noise. Yet this is interesting for car manufacturers in order to predict system performances for different speeds and different car models and thus allow to design speech based applications in a better way. It therefore makes sense to split the single WER into SNR dependent WERs, where SNR stands for the signal to noise ratio, which is an appropriate measure for the noise level. In this paper a SNR measure based on the concept of the Articulation Index is developed, which allows the direct comparison with human recognition performance.

1. Introduction

Comparing the performance of state of the art machine speech recognisers with the one of humans, it is clear that humans show higher recognition rates (Lippmann, 1997). Some argue that this is due to the 'world knowledge' of humans. The 'world knowledge' can be represented by the ability to predict the next word given a row of preceding words. For machine speech recognisers this is done by means of language models that model the statistical relationship between words. For the English language, language models were achieved, that need 1.75 bits to predict the succeeding letter (Brown, 1992). This value is close to the corresponding values for humans, which leads to the conclusion that the gap between machine and human speech recognition performance may not be due to the 'world knowledge' but rather to an insufficient acoustic-phonetic modeling of machine recognisers. This modeling describes the relation between the speech signal and the pronunciation of words. Humans as well as machine speech recognisers extract features based on the short time power spectrum from the speech signal. In humans this spectral analysis is done on the basilar membrane (Zwicker and Feldkeller, 1967), which has thoroughly been investigated so far. The subsequent first 'recognition' process in the acoustic cortex of humans is fairly unknown. It is commonly assumed that here the recognition of phonemes or syllables takes place. To investigate the acoustic recognition performance of this layer, H. Fletcher conducted a large amount of experiments (Allen, 1994a). His works were aimed at improving the intelligibility of the telephony transmission system.

2. Fletcher's Theory

Fletcher tried to determine the human recognition rates for phonemes ('sounds'). It was known that this recognition rate would depend on the context of the phonemes, whether the phonemes were spoken within meaningful context or within a non-sense context. Fletcher called the recognition rate for meaningful speech units like words or sentences 'Intelligibility'. He further called the recognition rate for speech units without context 'Articulation' ('sound articulation' s, 'syllable articulation' S). Fletcher investigated the Articulation by means of non-sense syllables. Most of the investigations were made with 80 CVC (Consonant-Vowel-Consonant) syllables placed at the end of short sentences. With speaker listener pairs recognition rates for syllables S as well as vowels v and consonants c were experimentally determined by playing non-sense syllables in sentences like: 'these sounds are lon'.

2.1. Articulation Index A

J.Q. Stewart started investigations concerning the determination of human phoneme recognition rates $e = 1 - s$ for bandpass filtered signals in 1921 (Allen, 1994a). In the following e_k represents the error rate measured with a speech signal that was band-pass filtered with the k_{th} bandpass. J.Q. Stewart found the following relationship:

$$e = e_1 e_2 ... e_n \tag{1}$$

where e is the error rate when all band passes are in use. $e_1, e_2, ..., e_n$ are all ≤ 1, that means each band B_k decreases the error rate by factor of e_k. Fletcher was looking for an additive term for the recognition rate composed of the recognition rates of the different bands. From equation (1) he derived the so called Articulation Index A by equations (2) and (3).

$$log(1-s) = log(1-s_1) + ... + log(1-s_n);\ e_k = 1-s_k \tag{2}$$

$$A = -\frac{Q}{p} log_{10}(1-s);\ 0 \leq A \leq 1 \tag{3}$$

Fletcher defined the 'practice factor' p to be 1 for a 'normal' speaker listener pair. Speakers with articulation difficulties e.g. would lead to a value of < 1. Q is a constant representing an intrinsic property of the system that was experimentally determined in the following way: For optimal transmission conditions the measured Articulation Index for a speaker-listener pair with $p = 1$ should reach the value of $A = 1$. Experiments under such conditions showed, that the test speaker-listener group of $p = 1$ achieved a phoneme recognition rate of 0.985. Put into equation (3) this leads to $Q = 0.55$. Thus equation (3) yields to (4).

$$A = -\frac{0.55}{p} log_1 0(1-s);\ s = 1 - 10^{-\frac{Ap}{0.55}};\ e = 10^{-\frac{Ap}{0.55}} \tag{4}$$

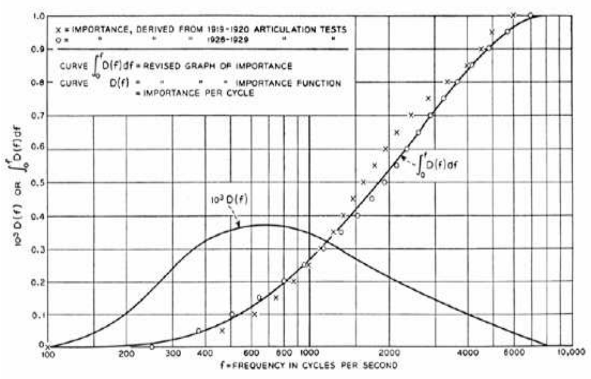

Figure 1: *Importance function D and Articulation Index A_f*

2.2. Importance Function D

In order to investigate the behaviour of the Articulation Index over frequency, the following experiment was conducted: A was investigated for a low-pass filtered signal. The cutoff frequency of the low-pass filter was gradually increased leading to a specific increase in A. From these experiments, which were conducted under optimal noise conditions, the importance function $D(f)$ was obtained:

$$D(f) = dA/df;\ A_f = \int_0^f D(f)df \qquad (5)$$

$D(f)$ can be interpreted as an information density. Fig. 1 (s.(Allen, 1994a) p.289) shows that D has a maximum at 700 Hz. That means the information density is highest here. Comparing $A_f(f)$ with the relationship between frequency and Bark scale (Zwicker and Feldkeller, 1967) it was found that the curves are very similar in the frequency range from 400Hz-5000Hz.

From equation (4) together with the $D(f)$ function, the Articulation Index can be written for arbitrary band limits according to:

$$e = 10^{-\frac{Ap}{0.55}};\ A = \int_{f_l}^{f_u} D(f)df \qquad (6)$$

where f_l and f_u are the lower and upper limits of a frequency band. From (6) the error rates for band-pass filtered un-noisy speech can be predicted. Using Fig. 1 the frequency scale from 0 Hz to 10 000 Hz can be divided into K_u frequency bands B_k of equal Articulation A_k. Fletcher calls these frequency bands with equal Articulation 'Articulation bands'. They have the following properties:

$$A_k = \int_{f_k-1}^{f_k} D(f)df;\ A_1 = ... = A_k; \qquad (7)$$

$$\sum_{k=1}^{K_u} A_k = 1 \rightarrow A_k = \frac{1}{K_u};\ k = 1, ..., K_u \qquad (8)$$

(Fletcher chose K_u = 20 as the number of Articulation bands). Together with equation (6) this leads to the same error rates for all Articulation bands:

$$e_k = 10^{-\frac{p}{0.55K_u}};\ k = 1, ..., K_u \qquad (9)$$

2.3. Modeling of noise

Up to now all investigations were made under ideal noise conditions (no noise). In the following the modeling of noise should be introduced to equation (6). It was shown that a given information density D for a frequency interval is reduced by noise by the factor of D_N. D_N is frequency dependent and is determined by the frequency dependent signal to noise ratio $SNR(f)$. Fletcher found in extension to (6) the following relationship:

$$e = 10^{-\frac{Ap}{0.55}};\ A = \int_{f_l}^{f_u} D_N(f)D(f)df \qquad (10)$$

where

$$D_N(f) = \begin{cases} SNR(f)/30\ for\ 0 \leq SNR(f) \leq 30dB \\ 1\ for\ SNR(f) > 30dB \\ 0\ for\ SNR(f) < 0dB \end{cases} \qquad (11)$$

Looking at the Articulation bands again, where the error rate of band k is termed e_k leads to:

$$e = \prod_{k=1}^{K_u} e_k = \prod_{k=1}^{K_u} 10^{-\frac{pD_{Nk}}{0.55K_u}} = 10^{-\frac{pD_N}{0.55K_u}} \qquad (12)$$

where $D_N = \sum_{k=1}^{K_u} D_{Nk};\ 0 \leq D_{Nk} \leq 1$

The values for D_{Nk} are determined by the mean SNR value of band B_k according to (11), and D_N characterises the noise of the whole frequency range. If Articulation bands are missing in a transmission system as is the case in telephony systems which are band limited, the respective D_{Nk} are set to 0.

3. Application of Fletcher's theory to an automatic speech recogniser

Especially for mobile applications noise conditions from the speaker environment are a major problem for machine speech recognisers. The noise conditions are mainly described by the signal to noise ratio (SNR). Investigating the relationship between recognition rates and SNR, first a measurement for the SNR is needed. This is not straight forward because during speech parts the noise is superposed to the speech signal and cannot be measured directly. Therefore an estimation of the SNR is needed. In the following, an SNR estimation method is introduced which is adapted to the properties of automatic speech recognisers. With this method, error rates are measured as functions of SNR and the results are related to equation (12).

3.1. Measuring the signal to noise ratio

State of the art speech recognisers usually make a 'Mel-cepstrum' analysis in order to extract appropriate features. Energy values of the so called Mel filters are calculated from the short-time power spectrum. This is done for short time sequences (10-30ms) the so called frames. The Mel filters are approximated band-passes with equal bandwidths on the Bark scale. The automatic speech recogniser under investigation has a bandwidth of 1.33 Bark for the Mel filters. The bandwidths of the Mel filters are comparable to the bandwidths of Fletcher's Articulation bands, at least in the frequency range from 400Hz-5000Hz (s. Fig.5 in (Allen, 1994b)). Therefore the SNR values will be calculated on the different Mel filters. The mean SNR value of the SNR values in the Mel filters will be called SNR_{Mel} in the following. The error rates will be given as a function of the SNR_{Mel}. To measure SNR_{Mel} a method had to be developed. As described in (Kim and M.Rahim, 2004) the method of forced Viterbi is used to segment speech signals into speech and non-speech parts. Like that every frame of 10-20ms is labeled 'speech' or 'non-speech'. For each frame i the signal energy E_i is calculated. In the non-speech parts this leads to the noise signal energy N_i. For the speech parts it is assumed that the noise signal N and the speech signal S are statistically independent. The mean energy for the speech parts $\overline{X}$ thus represents the sum of the mean speech energy and the mean noise energy. It is further assumed that the noise is stationary which means that the noise energy will be the same within and without the speech parts. This leads to the following calculation of SNR_{Mel} for a given utterance:

$$\overline{X}_k = \frac{1}{F_S}\sum_{i=1}^{F_S} X_i;\ F_S:\ number\ of\ speech\ frames$$

$$\overline{N}_k = \frac{1}{F_{NS}}\sum_{i=1}^{F_{NS}} N_i;\ F_{NS}:\ number\ of\ nonspeech\ fr.$$

$$SNR_{Mel}(k) = \frac{log_{10}(\overline{S}_k)}{log_{10}(\overline{N}_k)} = \frac{log_{10}(\overline{X}_k - \overline{N}_k)}{log_{10}(\overline{N}_k)}$$

$$SNR_{Mel} = \frac{1}{K}\sum_{k=1}^{K} SNR_{Mel}(k) \qquad (13)$$

where K ($K \leq K_u$) is the number of filter banks of the recogniser.

For the evaluation of this method clean speech files of the PhonDat database (web site, 2006) were added with noise recorded in car. This was done as described in (Höge et al., 2004). Like that, the clean speech files as well as the noise files were there to calculate the reference SNR_{Melref}. The SNR_{Mel} was then estimated on the mixed data with the method described in (13) and compared to SNR_{Melref}. It was found that the estimated Noise energies were too high compared to the reference noise energies. This was related to insufficiencies of the forced Viterbi algorithm which sometimes led to the inclusion of high energy speech parts into the noise estimation. This effect could be reduced by introducing a so called hangover. That means that a number of frames (here 10 frames) after a speech part are not used for the calculation of the signal energies. Furthermore outliers for the noise estimation were removed from the noise estimation in the following manner: First a statistic over the noise energies N_i for frames i for an utterance is made. The mean noise energy is then calculated with all noise frames N_i with: $\frac{N_i - mean(N_i)}{stdev(N_i)} < 2.7$. It was shown that the SNR_{Mel} thus calculated showed next to no difference to the reference SNR_{Mel}. In the following the SNR_{Mel} will always be calculated in the way described above.

3.2. Automatic versus Human Recognition rates

The investigations were done with an automatic speech recogniser based on Mel cepstral feature analysis and continuous mixture Gaussian density HMMs (Bauer, 2001). The HMM consisting of 1200 densities was trained using whole word modeling. For training and testing the German SpeechDat Car database (web site, 2006) was used. Figure 2 shows the results for a continuous spelling task. The results are shown with and without noise reduction. The noise reduction consists of a recursive least squares Wiener Filter (RLS). It increases the SNR by 10-15dB expanding the curve to higher SNR values.

In order to compare these results to Fletcher's theory, the human error rate according to equation (12) has to be matched with the error rate shown in Figure 2. The following steps are performed:

- It is assumed that the recognition of continuous spelling is similar to the recognition of phonemes in nonsense syllables.
- D_n in equation (12) is matched to the measured SNR_{Mel} from (13), which shows on the abscissa of Figure 2. As desribed above, critical bands and Melfilter bands share the same properties at least in the speech band. Equation (14) then yields the relationship between D_n and SNR_{Mel}.

$$\frac{D_N}{K_u} = \frac{1}{K_u}\sum_{k=1}^{K_u} D_k = \frac{K}{K_u}\frac{\sum_{k=1}^{K} SNR_{Mel}(k)}{K30} \qquad (14)$$

leading to:

$$D_N = \begin{cases} \frac{K}{K_u} SNR_{Mel}/30\ for\ 0 \leq SNR_{Mel} \leq 30dB \\ \frac{K}{K_u}\ for\ SNR_{Mel} > 30dB \\ 0\ for\ SNR_{Mel} < 0dB \end{cases} \qquad (15)$$

The investigations with the spelling recogniser were done with a band-limit of $f_g = 4000Hz$ corresponding to a number of bands K. Since every band taken into account has the same information, the ratio of number of bands K/K_u equals the ratio of information contained up to the respective bands. According to Fig. 1 this leads to a value K/K_u = 0.8. The human phoneme recognition error rate is thus determined by equation (16)

$$e = 10^{-\frac{pD_N}{QK_u}} \qquad (16)$$

From equations (16) and (15) the curve of the human error rate as a function of SNR_{Mel} is achieved (s. Figure 2).

3.3. Mapping Human onto ASR Error Rate

The similarity of the shape of the curves for the human and the automatic speech recognition leads to the idea of fitting the human curve onto the ASR curve and by doing that to obtain a measure for the quality of the ASR related to the one of humans.

Comparing the curves for human and for ASR error rates in Fig. 2, the following things can be observed:

- The SNR value where the machine recogniser reaches 100% error rate is higher than the 0 dB of the human recogniser.

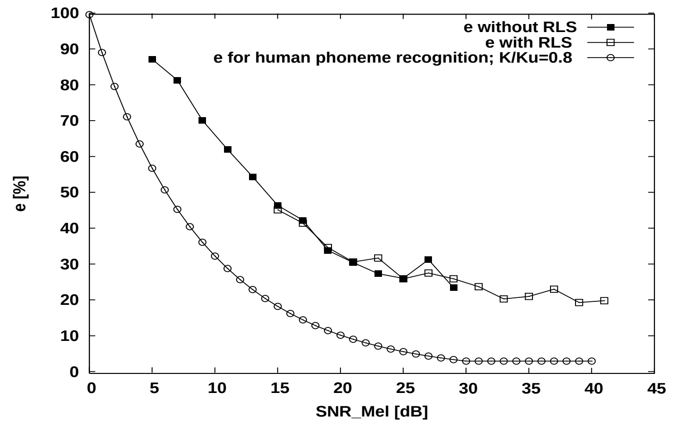

Figure 2: *Error rates if a continuous spelling task as a function of SNR_{Mel} with and without noise reduction*

- It is not clear if the machine recogniser reaches its maximum recognition rate at an SNR value of 30 dB as does the human recogniser.

- The error rate of the machine recogniser especially for high SNR is much worse than the error rate of the human recogniser.

One step to move the human curve towards the ASR curve is to change Q in equation (16). Q controls the minimum error rate for ideal noise conditions (s. 2.1.) and was found to be 0.55 for human phoneme recognition. From Fig. 3 the behaviour can be seen if e.g. Q is changed from 0.55 to 0.8.

Two further parameters have to be introduced to model the ASR curve out of the human curve. These parameters are SNR_u and SNR_l. SNR_u is the SNR value above which D_N is at a constant maximum. This value was found to be 30 dB for human beings (s. equation (11)). Likewise SNR_l is the SNR value under which D_N is at a constant minimum. This value was found to be 0 dB for human beings (s. equation (11)).

Equation (15) for D_N is thus modified yielding equation (17).

$$\frac{D_N}{K_u} = \begin{cases} \frac{K}{K_u} \frac{SNR_{Mel}-SNR_l}{SNR_u-SNR_l} \\ for\ SNR_l \leq SNR_{Mel} \leq SNR_u \\ \frac{K}{K_u}\ for\ SNR_{Mel} > SNR_u \\ 0\ for\ SNR_{Mel} < SNR_l \end{cases} \tag{17}$$

With values $SNR_l = 3dB$, $SNR_u = 27dB$ and $Q = 1.2$ an approximation of the ASR curve is reached as can be seen form Fig.3. The values of these three parameters allow an evaluation of the acoustic recognition performance of an automatic speech recogniser in comparison to the respective human recognition performance. Q and SNR_u represent the recognition performance for high and SNR_l for low SNR values. These three parameters appear to be sufficient to describe the recognition performance of a specific ASR over the whole SNR range. It will have to be investigated how these values behave with different HMMs. Also it will have to be investigated how these findings can be extended to the recognition of words.

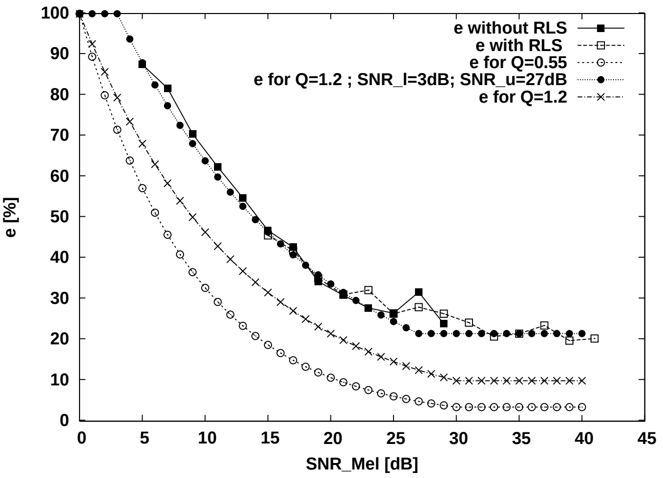

Figure 3: *Error rate 'e' of the spelling recogniser versus the human recognition rate a function of SNR_{Mel}*

4. Conclusion

In this paper the evaluation of in-car ASR was investigated. Instead of a single error rate the error rate was plotted as a function of the SNR. Since there is no standard SNR measurement, an SNR-measure based on the concept of the Articulation Index and the critical bands was developed. This allowed a direct comparison between human recognition performance and the one of a specific ASR. The comparison yielded a set of parameters, which can be used to assess the quality of ASR in comparison to human recognition performance.

5. References

J.B. Allen. 1994a. *Harvey Fletcher 1884-1981, The ASA Reprint of Speech and Hearing Communication*. J.B. Allen, Ed.:Acoustical Society of America, New York.

John B. Allen. 1994b. How do humans process and recognize speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577.

Josef G. Bauer. 2001. *Diskriminative Methoden zur automatischen Spracherkennung für Telefon-Anwendungen.* Ph.D. thesis, Technische Universität München.

P. F. Brown. 1992. An estimate of an upper bound for the entropy of english. *Computer Linguistics*, 18:31–40.

H. Höge, Josef G. Bauer, Christian Geissler, Panji Setiawan, and Kai Steinert. 2004. Evaluation of microphone array front-ends for asr - an extension of the aurora framework. *Proc. Fourth International Conference on Language Resources and Evaluations (LREC2004)*.

H.K. Kim and M.Rahim. 2004. Why speech recognizers make errors? *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*.

R. P. Lippmann. 1997. Speech recognition by machines and humans. *Speech Communication*, pages 1–15.

ELRA web site. 2006. http://www.elra.info.

E. Zwicker and R. Feldkeller. 1967. *das menschliche Ohr als Nachrichtenempfänger*. Kirzel Verlag, Stuttgart.