Predicting Word Correct Rate from Acoustic and Linguistic Confusability

Gies Bouwman, Bert Cranen, and Lou Boves

Radboud University Nijmegen, The Netherlands
{G.Bouwman,B.Cranen,L.Boves}@let.kun.nl

Abstract

When adapting an existing ASR-application for different user environments, one often gets confronted with speech that does not entirely match the training situation. Differences may stem both from acoustic and linguistic causes. In this paper we explore to what extent the word correct rate (WCR) for a given test set can be predicted from the transcription only (i.e. the linguistic representation) under the assumption that acoustic conditions are matched. We hope that, eventually, such a prediction can provide an estimate of a lower bound on WER to aim for when applying acoustic enhancement techniques. We propose and compute measures for *acoustic and linguistic confusability* (AC and LC) of each entry in the vocabulary of an ASR engine. Using a tabulation of how correctness of actual recognition on a development set varies as a function of these confusability measures, we show that actually observed WCR of words from independent test sets can be predicted with high accuracy over the full ranges of AC and LC levels.

1. Introduction

Automatic speech recognition assumes that speech units e.g. phones, words or other units, can be described as distinct and relatively invariant patterns. However, as soon as the testing conditions start to deviate from the conditions during training, the number of recognition errors tends to increase rapidly. The reason for this is that the ASR engine may misinterpret variation due to mismatch as systematic, speech related variation that it learned from the training material.

There may be many different reasons why the word error rate (WER) of a given test set is undesirably high. A well-known cause is the presence of background noise, which may introduce an acoustic training-test mismatch. Furthermore, speakers may have spoken unusually fast or slow, or they may have used less predictable word sequences (higher perplexity of the test set) or maybe the proportion of words that are easily confused acoustically was relatively high. All these factors interact with each other, so that in practice it is often very difficult to make out which factor constitutes the main contribution to the WER [1],[2]. In this paper we attempt to develop a diagnostic instrument that may be of help in unraveling the most important error sources.

The question that we will try to answer is whether it is possible to predict the word correct rate (WCR) of a given test set, when only the orthographic transcriptions of the utterances are used. This idea is not entirely new. For instance, in [2] a phoneme confusion model was used from which WER could be predicted for new ASR applications with new vocabularies and language models. With that purpose in mind, the authors decided not to use the acoustic models that were derived from the training data to avoid the possible drawback that wrong assumptions in the acoustic models might lead to under-estimation of the actual confusability.

In the current study our motivation is different. We would like to be able to obtain insight in the question to what extent the recognition performance is determined by linguistic factors under the assumption that acoustic characteristics of the set match with the training conditions and that the same acoustic models are used as during training. By doing so, we hope to automatically obtain an estimate of the maximum improvement that can be obtained when it appears necessary to apply noise reduction and normalization techniques because the test set is acoustically unmatched.

A method that would be suited for this purpose is described in [1]. In that approach acoustic observations are synthesized in order to compute the so-called *Synthetic Acoustic Word Error Rate*. However, because we eventually would like to investigate a priori WCR from the recognizer models point of view, we decided to try and develop a method that uses acoustic model parameters directly.

This paper has the following structure. Section 2 outlines the general procedure we propose. Sections 3 and 4 elaborate on the computation of acoustic and linguistic confusability. In Section 5 we describe the experimental set-up, followed by a presentation of the results in Section 6. Finally, Section 7 discusses the relation of the results with our original questions and provides our conclusions.

2. General method

As stated above, the aim of this paper is to develop and validate a method to predict WCR given only the transcription of a test corpus that has been recorded in similar acoustic conditions as the *train* corpus. In doing so, we will make two assumptions. First we will assume that the majority of the recognition errors are caused by the *intrinsic word confusability* and are in fact word substitution errors (i.e., not insertion or deletion errors). As discussed in the introduction, both language and acoustic models contribute to the degree of confusability of a word. Our second assumption is that WCR is basically determined by two variables that we will call acoustic confusability (AC) and linguistic confusability (LC).

We define a word w to be acoustically confusable if typical feature representations x_w encoding w will cause acoustic models of words $v\,(v \neq w)$ to produce likelihood scores similar to or higher than the score of w's own model. Similarly, a word w is said to be linguistically confusable if the (typical) linguistic reference context of w, like the N-1 preceding words in case of an N-gram, will induce the language model to assign similar or higher likelihood scores to words $v\,(v \neq w)$.

In the following sections we propose measures to define acoustic confusability and linguistic confusability of a test set on the basis of confusability of individual words. We compute an AC and LC score for each word in the transcription of a

development corpus with matching acoustic conditions. After conducting a recognition experiment, we will determine for each transcribed word whether it has been recognized correctly or has been substituted.

For all words with a similar {AC, LC}-score (we will subdivide the space of LC and AC values in 6×6 subspaces for this purpose) the average WCR is calculated and used as a predictor for the correct rate. Next, by repeating the procedure for a number of independent test sets of varying size, we investigate to what extent the predicted WCR differs in a statistically significant way from the observed correct rate.

We will design and validate our procedure by means of speech corpora that have been collected to develop and test the Dutch timetable information system VIOS [4]. The proportion of deletion and insertion errors is relatively small and will not be considered any further (in the VIOS corpus [4] we found that the major part (87 %) of all misrecognized words were substituted).

3. Acoustic confusability

Measures of *acoustic confusability* (AC) must express how severely realizations of a word are prone to substitution errors. One assessment strategy of AC would be to collect thousands of realizations of each lexicon entry and recognize them all with an ergodic isolated word grammar and count how often they are confused. However for most situations these amounts of data are not available.

A more practical alternative is to use a Kullback-Leibler (KL) divergence measure that essentially expresses a cross-entropy distance between two probability density functions (pdfs) cf. [3] and Eq. (1).

$$\mathrm{K}(d_v, d_w) = \sum_{n=1}^{N} \int \left[d_{nw}(x) - d_{nv}(x) \right] \cdot \log \left(\frac{d_{nw}(x)}{d_{nv}(x)} \right) dx \qquad (1)$$

where d_{nw} and d_{nv} represent the pdf's of the N individual vector components of the speech units v and w, e.g. HMM states, and the integration variable x covers the value domain of the n^{th} feature.

This divergence measure may be interpreted as follows: we use the acoustic model of one word to generate acoustic samples according to its density function, and the other to evaluate the 'costs' of each sample. Divergence values close to zero correspond to similar, i.e. confusable, acoustic feature distributions and vice versa. Thus, the KL divergence accumulated over all states of a word, can express acoustic confusability of v and w. This interpretation is compatible with the proposal by Printz and Olsen [1], but avoids the need to create an indirect representation of the generator models through a (vast) amount of synthesized tokens .

Since our recognizer models words as phoneme strings and phonemes by multi-state HMMs, each word is essentially represented by a chain of pdfs. To determine whether a pair of words is mutually confusable, we need to find a state-level alignment that corresponds to the alignment when acoustic feature vectors would have been there. To this end, we have implemented a DP search that finds the lowest KL divergence of all possible alignments. In order to allow for words to start and end at different points in time, a single silence state was pre-pended and appended to both state chains. The alignment procedure permitted self-loops, but states could not be skipped. Figure 1 presents a visualization of the search space and a path that corresponds to an alignment of the words 'beek' (SAMPA: /be:k/) and 'sneek' (/sne:k/). In this example, the path traverses the first HMM of 'sneek' before it enters the models of 'beek', showing different temporal starts of both words.

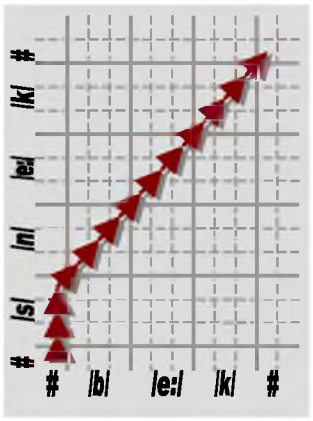

Figure 1 Word pair alignment conditioned by minimal Kullback-Leibler divergence.

In this way we obtained a 'minimal' accumulated divergence score for each pair of words, which is equal to the sum of all state divergence scores along the cheapest path. Ultimately, we needed a single score for each individual word w to represent its 'prior' confusability with *any* word v ($v \neq w$). During informal (unreported) analyses of previous recognition results, we had observed that words were commonly confused with only a limited number of other words. This observation led us to the idea to create a cohort set consisting of only the M nearest words of w. A score for this cohort was computed by taking the log of the average of the M exponentiated divergence scores.

The following is an excerpt of the verbose output of our program for cohort score computation from the top M=10 most confusable words for the station name 'Maarn'. The word is assigned a final cohort score of -2.78, which can be interpreted as 'quite confusable'. In the experiments reported here, we always used M=10.

COHORT SET for **w = maarn**

rk	<v>	<div(w,v)>	<e^div>
1.	baarn	-0.99	0.37
2.	maarssen	-2.52	0.08
3.	naarden	-2.88	0.06
..	...	...	...
10.	hoorn	-4.47	0.01
AC SCORE:		-2.78	0.06

4. Linguistic confusability

As defined in Section 2, we want the measure of *linguistic confusability* of a transcribed word w to express the entropy [i.e., (lack of) predictive power] of the Language Model (LM) when it renders the prior probability for w during search.

Since our continuous speech recognizer uses N-gram LMs, our LM was designed to model the prior probability of the i-th word of a sentence w_i as

$$p(w_{i-N+1} \ldots w_i) = p(w_i \mid h_i^w) \qquad (2)$$

where $h_i^w = (w_{i-N+1} \ldots w_{i-1})$ denotes the history of w_i constrained by i-N+1 ≥ 0 and w_0 equals the representation for the start of an utterance.

We distinguish two types of LC and we wish to compare them in our experiments. The first (LM-based LC, $\mathrm{LC}^{\mathrm{LM}}$) is computed for each lexicon entry using the Language Model *irrespective of the word's N-gram context in the transcription*. The other (transcription-based LC, $\mathrm{LC}^{\mathrm{transc}}$)

does take the N-gram history of each word in the transcription into account, and makes full use of the local linguistic context in which the word was realized.

4.1. LM-based LC

This measure represents LC of w as the weighted average of the LM entropies for the top M most frequent history contexts for w [cf. Eq. (5)]. "Most frequent" refers to observation counts in the LM training corpus. The contribution weights of these M entropy values to the average are linearly proportional to the frequency.

The top M most frequent N-gram histories of w can be found by computing

$$H_M(w) = \left\{ \arg\max_{h}{}^{M-\text{fold}} \left(p(w|h) \cdot p(h) \right) \right\} \quad (3)$$

$H_M(w)$ will be abbreviated as H hereafter. The relative frequency that each of these histories h_j ($j = 1...M$) occurs in combination with w is represented by the normalised prior probability:

$$p_{norm}(w, h_j) = \frac{p(w|h_j) \cdot p(h_j)}{\sum_{g \in H} p(w|g) \cdot p(g)} \quad (4)$$

Now, we define LM-based LC of w as

$$LC^{LM}(w) = \sum_{g \in H} \left(p_{norm}(w, g) \cdot {}^2\log\, p(w|g) \right) \quad (5)$$

In all of our experiments we used N=2 (bigram) and M=20. The following is an excerpt of $H_{20}(aan)$: the top twenty bigram histories (single words) that precede the word 'aan'.

HISTORYSET for **w = aan**

rk	<h>	$<^2\log p(w\|h)p(h)>$	$<p_{norm}>$	$<^2\log p(w\|h)>$
1.	alphen	-10.87	0.8022	-0.0751
2.	voldoende	-13.96	0.0938	-3.6663
3.	capelle	-15.04	0.0444	-1.6429
..	...	...	...	...
20.	komen	-20.05	0.0014	-12.3730

LC SCORE: -1.09 (perplexity = 2.13)

In this example the bigram score of the word 'alphen' is largely responsible for the low total LC^{LM} score (-1.09 → low confusability). Largely responsible, because it occurs in 80.2% of the top 20 bigrams of which 'aan' is the second word. Low LC^{LM}, because 'alphen' is hardly ever followed by a word ≠ 'aan' for the simple reason that 'alphen' is the fir st part of the city name 'Alphen aan de Rijn'. In summary, the word 'aan' is marked as *linguistically not confusable*, because the LM will give it a high prior probability in the majority of the linguistic contexts where it is spoken.

4.2. Transcription-based LC

As mentioned before, we assume a transcription to be available, for we make a predictor of WCR based on the transcription. Our alternative LC, called LC^{transc}, incorporates the transcription of the full N-gram context in which each individual word w_i was spoken:

$$LC^{transc}(w_i) = {}^2\log\, p(w_i \mid h_i^W) \quad (6)$$

Of course, although the arguments w_i and h_i^W of the N-gram probability are derived from the transcription, the probability $p(w_i \mid h_i^W)$ itself is independently estimated on the training corpus of the language model. This measure expresses the linguistic confusability induced by h_i^W. In fact, averaging the negative of LC^{transc} over all words of the transcription and taking that in the power of 2 is commonly known as the testset perplexity. In all experiments reported here, we used N=2 (bigram).

5. Experimental set-up

Having discussed the methods to provide each of the words in the transcription with an acoustic and linguistic confusability score, we will now present the remaining properties of our experimental set-up.

The VIOS speech recordings were stored as a-law audio files and orthographically transcribed. Our data were divided over three non-overlapping corpora:

- **training corpus** 10h = 84k words, used for training HMMs and LM
- **development corpus** 2.5h = 20k words, used for assessing the numerical relation (AC, LC)→WCR
- **test corpus** 22h = 189k words, used to verify predictive capabilities. This corpus was randomly partitioned at four levels T1...T4 with respectively 1×189k (full set), 9×20k, 18×10k and 37×5k words.

The unusual choice for a larger test corpus than the training corpus is due to historic reasons related to incremental availability of data.

For our experiments we used the PHICOS recognition engine (as in [4]). Feature vectors were extracted from 16 ms time frames at 100 Hz. Each frame was represented by a vector containing log energy, 13 MFCCs and their deltas, i.e. 28 features. The monophone HMMs had a tristate left-to-right topology. States were modeled by mixture pdfs that consisted of at most 32 Gaussians, depending on availability of training samples. The lexicon contained 984 words, of which 436 are (parts of) station names. Finally, the language was described by a word bigram model, which was trained on the 84k bigrams of the afore-mentioned training corpus.

6. Results

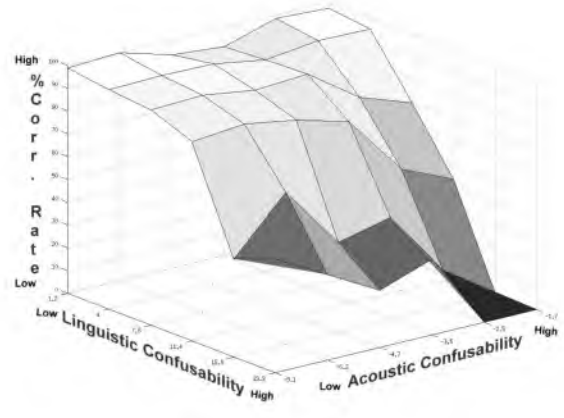

Figure 2 Word Correct Rate plane as a function of LC^{transc} and AC

6.1. CR on development set

Figure 2 illustrates the result for LC^{transc} and AC versus WCR in a 3-D plot, as determined on the development set. Each intersection of grid lines on the plane represents the corresponding average WCR. Analogously, we computed the numerical relation of LC^{LM} and AC with WCR.

6.2. Validation of predictive capability

Having registered the expected WCR for the 6×6 range clusters, we compare them with actual WCR values of the test sets. To this end we computed z-values of statistical differences between proportions (5% level, two-sided). Figure 3 shows greyscale images corresponding to an example with a 20k word test set. The z-value of 2.16 (lightest square, right panel) indicates the only significant ($|z| > 1.96$) difference. In this case, the observed WCR of 95.5% (based on 870 tokens, corresponding cell in left panel) exceeded the predicted WCR of 92.1% by 2.4% (middle panel).

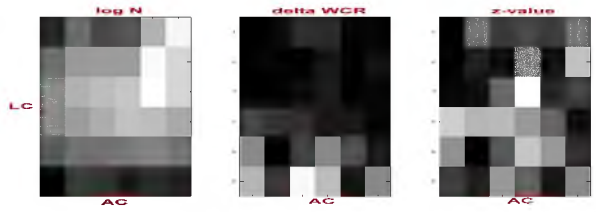

Figure 3 Three greyscale images of log #tokens, ΔWCR and z-values. (white=large; black=small)

For all test set sizes (T1…T4) we saw cluster WCRs that differed significantly from the prediction, although we did not observe a systematic pattern with respect to the (AC,LC)-positions of these clusters.

6.3. Summarization

Table 1 shows percentages of (cluster) points where the actual test set WCR did *not* differ statistically from the predicted WCR. Parenthesized are minimum and maximum percentage.

Table 1 Average % of clusters having no statistical difference between WCR prediction and reality.

set	# words	$AC+LC^{LM}$	$AC+LC^{transc}$
T1	1×189k	89% (-)	94% (-)
T2	9×20k	92.4 (91-100)	94.7 (92-100)
T3	18×10k	92.4 (91-100)	95.3 (92-100)
T4	37× 5k	89.6 (85-100)	94.6 (89-100)

7. Discussion and conclusions

The main question we wanted to answer in this paper is whether average word correct rate can be predicted based on a transcription of a test set for which acoustic conditions match the training conditions. Our results indicate it can; statistical agreement among prediction and observation was found for 89% or more of the AC, LC clusters in all test conditions.

When considering the columns in Table 1, the third column displays lower agreement values than the fourth. As expected, the LC based on the transcription is more accurate. Just like the recognizer, it takes the predictive ability of the immediate context - i.e. the preceding word - into account.

When comparing the rows (T1…T4), a clear tendency does not seem to be present, except for the minimum agreement percentage: for the T4 sets, the lowest agreement was found to be 85%. Until closer analyses point out otherwise, we believe that we may suffer from idiosyncrasies due to limited data here (5k words spread out over 36 points).

As far as Figure 2 is concerned, we would like to draw attention to the shape of the 'WCR landscape' as a function of AC and LC. We observe that expected accuracy is quite high if LC is low, irrespective of the AC level. However, as LC increases, we sooner or later cross the edge of an abyss, after which the word correct rate seems to collapse all of a sudden. If AC is below a specific level, this moment is delayed. This observation corroborates the conclusions of various studies on lexical modeling of pronunciation variation (e.g. [5]), where it was found that extension of the lexicon with new variants saturates modeling power at some point, after which the acoustic confusability introduced by new variants starts to affect recognition performance negatively.

We proposed a method to compute the acoustic and linguistic confusability of the words of a corpus, given their transcription. For different corpora that were recorded under conditions similar to those of the training corpus, these confusability scores were shown to accurately predict actual word correct rates for sets of words with similar acoustic and linguistic confusability scores. From this we conclude that it is possible to predict WCR even when test speakers have adopted a use of language that implies the use of more confusable words or improbable word orders compared to the training conditions. As a consequence, the predicted WCR should also provide an estimate of the maximum performance that can be obtained in acoustic conditions that do not match the training situation.

Acknowledgement

The current research is part of the CRIMI project (`http://crimi.id.tue.nl`) and is funded by the Dutch Ministry of Economic Affairs through the Innovation Oriented Programme Man-Machine Interaction (IOP-MMI).

References

[1] Printz, H. and Olsen, P., "Theory and Practice of Acoustic Confusability". *Proc. ISCA ITRW ASR2000,* CD-ROM 2000, Paris.

[2] Deng, Y., Mahajan, M., and Acero, A., "Estimating Speech Recognition Error Rate without Acoustic Test Data". *Proc. Eurospeech 2003*, 2003, Geneva, pp. 929-932.

[3] De Wet, F., de Veth, J., Cranen, L., and Boves, L., "Accumulated Kullback Divergence for Analysis of ASR Performance in the Presence of Noise". *Proc. ICASSP*, CD-ROM, 2003, Hong-Kong.

[4] Strik, H., Russel, A., Heuvel, H. v.d., Cucchiarini, C., and Boves, L., "A spoken dialogue system for the Dutch public transport information service", Int. Journal for Speech Technology, Vol. 2, No. 2, 1997, pp. 119-129.

[5] Deligne, S., Maison, B., and Gopinath, R., "Automatic generation and selection of multiple pronunciations for dynamic vocabularies". *Proc. ICASSP*, CD-ROM, 2001, Salt Lake City.