EXPLICIT WORD ERROR MINIMIZATION USING WORD HYPOTHESIS POSTERIOR PROBABILITIES

Frank Wessel, Ralf Schlüter, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department,
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{wessel, schlueter, ney}@informatik.rwth-aachen.de

ABSTRACT

In this paper, we introduce a new concept, the *time frame error rate*. We show that this error rate is closely correlated with the *word error rate* and use it to overcome the mismatch between *Bayes' decision rule* which aims at minimizing the expected *sentence error rate* and the word error rate which is used to assess the performance of speech recognition systems. Based on the time frame errors we derive a new decision rule and show that the word error rate can be reduced consistently with it on various recognition tasks. All stochastic models are left completely unchanged. We present experimental results on five corpora, the Dutch *Arise* corpus, the German *Verbmobil* '98 corpus, the English *North American Business* '94 20k and 64k development corpora, and the English *Broadcast News* '96 corpus. The relative reduction of the word error rate ranges from 2.3% to 5.1%.

1. INTRODUCTION

Statistical decision theory aims at minimizing the expected cost of making errors. For speech recognition this cost is defined as the cost of choosing a sentence $v_1^M = v_1, \ldots, v_M$ instead of the presumably correct sentence $w_1^N = w_1, \ldots, w_N$. In this very general framework, it is left open how the cost $\mathcal{C}(w_1^N, v_1^M)$ is defined:

$$\{w_1^N\}_{\text{opt}} = \operatorname*{argmin}_{w_1^N} \Big\{ \sum_{v_1^M} \mathcal{C}(w_1^N, v_1^M) \cdot p(v_1^M | x_1^T) \Big\} \quad , \qquad (1)$$

where $p(v_1^M|x_1^T)$ is the posterior probability for sentence v_1^M, given the acoustic observations $x_1^T = x_1, \ldots, x_T$. The standard approach in statistical speech recognition is to use the sentence error rate (*SER*) as a cost function. With this simple cost function, Bayes' decision rule can easily be derived:

$$\begin{aligned} \{w_1^N\}_{\text{opt}} &= \operatorname*{argmax}_{w_1^N} \Big\{ p(w_1^N | x_1^T) \Big\} \\ &= \operatorname*{argmax}_{w_1^N} \Big\{ \frac{p(x_1^T|w_1^N) \cdot p(w_1^N)}{p(x_1^T)} \Big\} \quad , \end{aligned} \qquad (2)$$

where $p(w_1^N)$ denotes the *language model* probability, $p(x_1^T|w_1^N)$ the *acoustic model* probability, and $p(x_1^T)$ the probability of the acoustic observations. The advantage of this simple cost criterion is that the resulting decision rule can be evaluated quite easily. Unfortunately, we are now faced with a conceptual mismatch between the decision rule and the evaluation criterion for speech recognizers, the word error rate *(WER)*. The authors of [6] present an example which shows that minimizing the expected sentence error rate does not necessarily minimize the expected word error rate. The easiest way to overcome this mismatch is to use the same cost function for optimization as for evaluation, the *Levensthein* distance between two sentences w_1^N and v_1^M which counts the number of word deletions, insertions and substitutions. The main drawback of this approach is its computational complexity, since it requires the pairwise alignment of all possible sentences.

In [6], the pairwise alignment is therefore restricted to the sentences in an *N-best list*. The sentence posterior probabilities are also approximated on these lists. This N-best list based approximation of the posterior probabilities was previously used in [7] for keyword spotting and in [9, 11] to compute confidence measures. The authors of [7] report a reduction of the word error rate by 1.0% relative on the *Switchboard* and the *CallHome Spanish* corpus.

In [3], *word graphs* are used instead of N-best lists. Unfortunately, an explicit Levensthein alignment of all pairs of sentences contained in the graph is prohibitive, since word graphs contain a considerably higher number of sentences. The non-local Levensthein alignment (in the sense that there is no straight-forward factorization which could be used to compute an alignment for all sentences at the same time) is therefore replaced by a multiple string alignment. With this multiple alignment, the posterior probabilities can easily be approximated and a simple decision rule can be derived. The authors report a reduction of the word error rate by 3.6% relative on the *Switchboard* corpus.

2. TIME FRAME ERRORS

As already discussed, the main problem of an explicit alignment of all sentences in a word graph is the non-locality of the dynamic programming alignment which is caused by deletions and insertions. Would the situation change, if we had no deletions and insertions? Let us assume that substitutions were the only type of error. In this case, all sentences would be of equal length and the positions of the words in the sentences would already define a multiple alignment. A dynamic programming alignment would thus not be necessary.

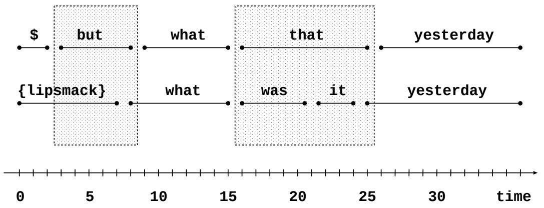

Figure 1: Illustration of the time frame error rate concept. \$ denotes a silence and {lipsmack} a noise hypothesis. The shaded boxes illustrate where time frame errors occur.

With these considerations in mind and with the fact that a word graph contains the starting and ending times of the word hypotheses, we study a new cost function that we decided to denote as time frame error rate. Before going into details, we first define the term word graph. In this paper, a word graph is a directed, acyclic, weighted graph. Its nodes represent discrete points in time, its edges word hypotheses $[w;\tau,t]$ for word w from node τ to node t, and its weights the acoustic probabilities of the hypotheses. Let $t_1^N = t_1, \ldots, t_N$ denote the ending times of a sequence of word hypotheses w_1^N. The starting time for word w_n is thus given as $t_{n-1}+1$, where $t_0 = 0$ and $t_N = T$. Each path through the word graph is a sentence hypothesis and can be written as $[w;t]_1^N = [w_1; t_0+1, t_1], \ldots, [w; t_{N-1}+1, t_N]$. We can now rewrite Eq. (1) so that the minimization and the summation are also over the unknown starting and ending times of the words. The assumption we make is that the cost function now also depends on the word boundaries and that the word graph as a limited representation of the search space contains enough sentence hypotheses.

2.1. Definition of the Time Frame Errors

We will now specify the notion of time frame error. Consider two sequences of word hypotheses contained in the word graph, $[w;t]_1^N$ and $[v;\tau]_1^M$. For each point in time $\hat{t}$ we can evaluate whether the word identities of the hypotheses in both sequences for time frame $\hat{t}$ are identical or not. Fig. 1 illustrates this concept. The shaded boxes indicate where time frame errors occur. The time frame errors are caused either by word deletions, insertions, and substitutions or by differing word boundaries. As Fig. 1 also shows, a substitution of lexical entries which are not counted as word errors (e.g. hesitations, noises, and silence) with each other is also not counted as a time frame error. We can now define a new cost function using the notion of time frame errors:

$$\mathcal{C}([w;t]_1^N, [v;\tau]_1^M]) = \sum_{n=1}^{N} \frac{\displaystyle\sum_{\hat{t}=t_{n-1}+1}^{t_n} 1 - \delta(w_n, v_{\hat{t}})}{1 + \alpha \cdot (t_n - t_{n-1} - 1)} \quad , \tag{3}$$

where $v_{\hat{t}}$ is the word identity of the word hypothesis in sentence $[v;\tau]_1^M$ which intersects time frame $\hat{t}$. We should note that the quantities which are compared using the Kronecker function in Eq. (3) are generalized word labels. Substitutions of words with the same label (e.g. silence, hesitations, and noises) are not counted as errors, as illustrated in Fig. 1.

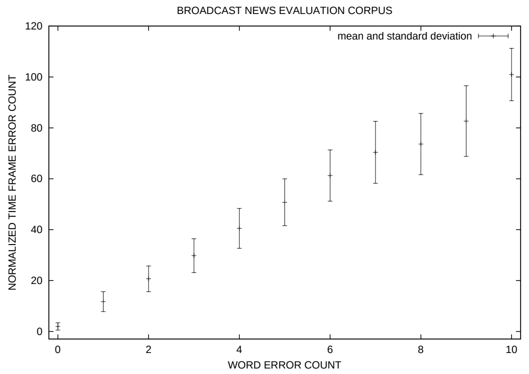

Figure 2: Plot of the mean and the standard deviation of the normalized time frame error counts for $\alpha = 0.05$ over the word error counts for the Broadcast News evaluation corpus. The plots for other values of α look very similar.

The denominator in Eq. (3) is used to normalize the time frame errors smoothly. For $\alpha = 0$ no normalization takes place, whereas for $\alpha = 1$ the time frame errors are fully normalized with the length of the current hypothesis. A possible explanation for the usefulness of the normalization is that it increases the effect of time frame errors for short word hypotheses. It is reasonable to argue that time frame errors are more "significant" if the current word hypothesis is very short, since short hypotheses tend to cause more word errors than longer word hypotheses. By choosing an appropriate α we can thus adjust the effect of short hypotheses on the total cost.

The main advantage of our new concept is that there are only substitutions on a time frame level. The words are either identical or not and no time consuming alignment of the hypotheses is necessary.

2.2. Correlation Analysis

It is obvious that for zero time frame errors we will also have zero word errors. For time frame errors above zero the word errors will also rise. The only question is, whether both error types are correlated, so that when minimizing one of them, the other one is also reduced. In order to study the correlation, we computed large N-best lists for our testing corpora (they are described later on) which also contained the starting and ending times of the word hypotheses. For each pair of sentences in the N-best list we then computed the Levensthein alignment, the word errors, and the time frame errors. Fig. 2 shows a plot of the mean and the standard deviation of the normalized time frame errors over the word errors per sentence on the Broadcast News testing corpus for $\alpha = 0.05$. The plots for the other corpora look very similar and are omitted due to the limited space.

In addition, we computed the correlation coefficients between the word errors and the time frame errors for $\alpha = 0.05$, see Table 2. As the experiments clearly show, there is a significant correlation between the word errors and the time frame errors. By reducing the time frame error rate we should thus be able to reduce the word error rate.

$$\left\{[w;t]_1^N\right\}_{\text{opt}} = \operatorname*{argmin}_{[w;t]_1^N} \left\{ \sum_{[v;\tau]_1^M} \mathcal{C}([w;t]_1^N,[v;\tau]_1^M) \cdot p([v;\tau]_1^M|x_1^T) \right\} \quad (4)$$

$$= \operatorname*{argmin}_{[w;t]_1^N} \left\{ \sum_{[v;\tau]_1^M} \sum_{n=1}^{N} \frac{\sum_{\hat{t}=t_{n-1}+1}^{t_n} 1-\delta(w_n,v_{\hat{t}})}{1+\alpha\cdot(t_n-t_{n-1}-1)} \cdot p([v;\tau]_1^M|x_1^T) \right\} \quad (5)$$

$$= \operatorname*{argmin}_{[w;t]_1^N} \left\{ \sum_{n=1}^{N} \frac{\sum_{\hat{t}=t_{n-1}+1}^{t_n} \left[1-\sum_{[v;\tau]_1^M} \delta(w_n,v_{\hat{t}})\cdot p([v;\tau]_1^M|x_1^T)\right]}{1+\alpha\cdot(t_n-t_{n-1}-1)} \right\} \quad (6)$$

$$= \operatorname*{argmin}_{[w;t]_1^N} \left\{ \sum_{n=1}^{N} \mathcal{S}([w_n;t_{n-1}+1,t_n]) \right\} \quad \text{, where} \quad \mathcal{S}([w_n;t_{n-1}+1,t_n]) = \frac{\sum_{\hat{t}=t_{n-1}+1}^{t_n} \left[1-p(w_n|\hat{t},x_1^T)\right]}{1+\alpha\cdot(t_n-t_{n-1}-1)} \quad (7)$$

2.3. Decision Rule

We can now insert the new cost function, cf. Eq. (3), into the general decision rule, cf. Eq. (4). After some simple manipulations we obtain our new decision rule, cf. Eq. (7). The new probability density function $p(w_n|\hat{t},x_1^T)$ can be interpreted as the probability to observe word w_n at time frame $\hat{t}$, given the acoustic observations:

$$p(w_n|\hat{t},x_1^T) = \sum_{[v;\tau]_1^M} \delta(w_n,v_{\hat{t}})\cdot p([v;\tau]_1^M|x_1^T) \quad (8)$$

$$= \sum_{[v;\tau]_1^M} \sum_{\substack{m=1: \\ \tau_{m-1}<\hat{t}\le\tau_m}}^{M} \delta(w_n,v_m)\cdot p([v;\tau]_1^M|x_1^T) \quad (9)$$

$$= \sum_{\substack{[v;\tau,t]: \\ \tau\le\hat{t}\le t}} \delta(w_n,v)\cdot p([v;\tau,t]|x_1^T) \quad (10)$$

As Eqs. (8) and (9) show, the sum is over the posterior probabilities of all sentences which contain a hypothesis for word w_n at time frame $\hat{t}$. Assuming that we have already computed posterior probabilities $p([v;\tau,t]|x_1^T)$ for individual word hypotheses, Eq. (9) can be rewritten so that the summation is over the posterior probabilities of all word hypotheses for word w_n which intersect time frame $\hat{t}$, cf. Eq. (10). These word hypothesis posterior probabilities are defined as the sum of the posterior probabilities of all paths passing through $[v;\tau,t]$ and can be computed very efficiently with a forward-backward algorithm on word graphs. In [8, 9, 11], these posterior probabilities were used to compute confidence measures. In [10], they were used to improve the recognition performance in the standard Bayesian approach. Regarding the posterior probabilities as the probability of a word being *correct*, the new hypothesis score $\mathcal{S}([w_n;t_{n-1}+1,t_n])$ can be interpreted as the normalized probability of a word being *incorrect*. The decision rule simply picks that sequence of word hypotheses with the minimum expected number of errors.

Table 1: Description of the testing corpora and the word graphs. WGD denotes the word graph density, GER the word graph error rate, and WER the word error rate.

corpus	size of vocab.	WGD	perpl.	GER [%]	WER [%]
Arise	985	218.8	12.6	7.4	15.8
Verbmobil	7128	209.2	56.1	8.7	33.6
NAB 20k	19987	98.4	124.5	4.1	13.2
NAB 64k	64736	87.1	145.9	1.8	11.1
BN	65491	105.5	213.7	10.6	33.3

3. SPEECH CORPORA

We first describe the five different speech corpora used for the correlation analysis and the rescoring experiments which will be presented later. The English North American Business (NAB) '94 development corpora consist of high-quality recordings of read newspaper articles. The Broadcast News '96 (BN) evaluation corpus consists of television and radio broadcast, whereas the German Verbmobil '98 evaluation corpus consists of spontaneous human-to-human dialogues, see [1]. The Dutch Arise corpus is composed of human-to-machine dialogues, recorded over the telephone with an automatic train timetable information system, see [4]. Table 1 specifies the experimental setup. The *word graph density* is defined as total number of word graph edges divided by the number of spoken words. The *graph error rate* is the lowest word error rate that can be achieved with a given word graph. For details the reader is referred to [5]. All word graphs were generated with our speech recognizer using gender independent acoustic models without speaker adaptation and a trigram language model.

4. RESCORING EXPERIMENTS

In order to simplify the search for the best path with regard to our new criterion, we compute the word hypothesis pos-

Table 2: Correlation coefficients for $\alpha = 0.05$ between the word errors and the time frame errors on a sentence level and results for the rescoring experiments with sentence error rate (SER) and time frame error rate (TFER) criterion.

corpus	correlation coefficient	SER criterion		TFER criterion	
		del - ins - WER [%]	SER [%]	del - ins - WER [%]	SER [%]
Arise	0.84	2.1 - 3.2 - 15.8	24.3	3.7 - 2.2 - 15.0	23.6
Verbmobil	0.93	6.1 - 6.9 - 33.6	82.9	8.0 - 5.2 - 32.5	83.6
NAB 20k	0.95	1.9 - 2.1 - 13.2	79.6	2.2 - 1.9 - 12.9	79.7
NAB 64k	0.95	2.0 - 1.5 - 11.1	74.8	1.8 - 1.6 - 10.8	75.8
BN	0.94	6.0 - 4.3 - 33.3	91.4	7.9 - 3.3 - 32.3	91.6

terior probabilities and then the new hypothesis score for each hypothesis in the word graph. The search algorithm is based on our standard word graph rescoring algorithm [5]. Instead of the acoustic scores stored during the generation of the word graph, we simply use the new hypothesis score. The language model is no longer needed at this stage.

Table 2 presents the results for the five testing corpora. As can be seen, the word error rates are reduced significantly. The relative reduction ranges between 2.3% and 5.1% and is highest for corpora consisting mainly of spontaneous speech. In order to obtain these results, we had to normalize the time frame errors for all corpora with an α in the range of $0.01 < \alpha < 0.1$. The effect of the specific choice of the normalization parameter in this range is very small. In order to rule out search errors during the standard word graph rescoring with the SER criterion which might explain the improved performance of the new TFER criterion, we applied no pruning during the rescoring process. We also verified that the performance of the SER criterion cannot be further improved with an additional word penalty.

As expected, the sentence error rates rise for the new criterion with the exception of the Arise corpus whose average sentence length is three words. For this corpus, the word error rate and the sentence error rate are stronger correlated than for the other corpora. In contrast to [6], we observed a significant reduction of the word error rate also for recognition tasks with rather low word error rates.

5. CONCLUSIONS AND OUTLOOK

In this paper, we presented a new cost function for speech recognition, the time frame error rate. Our experiments showed a significant correlation between the time frame errors and the word errors. Based on this new error rate we derived a criterion which directly aims at minimizing the time frame error rate and thus the word error rate instead of the sentence error rate. With the suggested method, the word error rates were reduced significantly on five different testing corpora. The relative reduction ranges between 2.3% and 5.1%.

In the future, we will investigate why an appropriate α is needed for normalization. In particular, we will study the word errors and time frame errors for individual words to find a more systematic explanation. Also, the new cost function could be used for the recognition of content words. In such a scenario, the generalized word labels discussed above could be used to map all words which are not relevant in this context to the same class so that they are not counted as errors. A similar approach is presented in [2].

6. REFERENCES

[1] T. Bub and J. Schwinn, "VERBMOBIL: The evolution of a complex large speech-to-speech translation system," in *Proc. ICASSP*, Philadelphia, PA, USA, Oct. 1996, pp. 2371–2374.

[2] V. Goel and W. Byrne, "Minimum Bayes Risc Automatic Speech Recognition," *Computer Speech and Language*, vol. 14, no. 2, pp. 115–135, 2000.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. EUROSPEECH*, Budapest, Hungary, Sept. 1999, pp. 495–498.

[4] J. Mariani and L. Lamel, "An overview of EU programs related to conversational/interactive systems," in *Proc. of the 1998 Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, Feb. 1998, pp. 247–253, Morgan Kaufman Publishers.

[5] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, Jan. 1997.

[6] A. Stolcke, Y. König, and M. Weintraub, "Explicit word error rate minimization in N-best list rescoring," in *Proc. EUROSPEECH*, Rhodes, Greece, Sept. 1997, pp. 163–166.

[7] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *Proc. ICASSP*, Detroit, MI, USA, May 1995, pp. 297–300.

[8] F. Wessel, K. Macherey, and R. Schlüter, "Using word probabilities as confidence measures," in *Proc. ICASSP*, Seattle, WA, USA, 1998, pp. 225–228.

[9] F. Wessel, K. Macherey, and H. Ney, "A comparison of word graph and N-best list based confidence measures," in *Proc. EUROSPEECH*, Budapest, Hungary, Sept. 1999, pp. 315–318.

[10] F. Wessel, R. Schlüter, and H. Ney, "Using posterior probabilities for improved speech recognition," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1587–1590.

[11] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, to be published, 2001.