



On the relation of Bayes Risk, Word Error, and Word Posteriors in ASR

Ralf Schlüter, Markus Nussbaum-Thom, and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany.

{schlueter,nussbaum,ney}@cs.rwth-aachen.de

Abstract

In automatic speech recognition, we are faced with a well-known inconsistency: *Bayes* decision rule is usually used to minimize sentence (word sequence) error, whereas in practice we want to minimize word error, which also is the usual evaluation measure. Recently, a number of speech recognition approaches to approximate *Bayes* decision rule with word error (*Levenshtein*/edit distance) cost were proposed. Nevertheless, experiments show that the decisions often remain the same and that the effect on the word error rate is limited, especially at low error rates. In this work, further analytic evidence for these observations is provided. A set of conditions is presented, for which *Bayes* decision rule with sentence and word error cost function leads to the same decisions. Furthermore, the case of word error cost is investigated and related to word posterior probabilities. The analytic results are verified experimentally on several large vocabulary speech recognition tasks.

1. Introduction

In automatic speech recognition research, a number of approximate approaches to word error minimizing decision rules were introduced recently¹ [1, 7, 9, 11], trying to overcome the well-known inconsistency between the standard decision rule optimizing sentence error and the appropriate word level evaluation measures [5, pp. 4-5].

Nevertheless, improvements obtained with these approaches usually are relatively small, where the larger improvements usually are obtained on more demanding tasks, with higher baseline error rates. Most important, experimentally it can be observed that the decisions using the standard sentence (word sequence) cost function often are equal to those obtained with decision rules based on word level (*Levenshtein* or similar) cost functions, as used in ASR, cf. e.g. the individual systems' results presented in [4].

In this work, we analyze the relation between the sentence (word sequence) cost function, and the *Levenshtein* cost function as should ideally be used in ASR. Specific properties for the word error minimizing *Bayes* decision rule, and its relation to a specific definition of word posterior probabilities will be derived. The analytic results are verified and evaluated experimentally. Their quantitative effect is studied by experiments on three different well-known large vocabulary ASR tasks.

2. Bayes Decision Rule

Assume the case of ASR, with acoustic feature vector sequences $X = x_1, \dots, x_T$, and word sequences $W = w_1, \dots, w_N$.

¹In automatic speech recognition, these approaches sometimes are denoted as "minimum" *Bayes* risk, although *Bayes* risk by definition already is minimal. The actual aim in these approaches is to apply word error cost functions to *Bayes* decision rule.

Also, define a cost function $\mathcal{L}(V, W)$. Then the posterior risk for a hypothesizing a word sequence W given an observation sequence X is

$$\mathcal{R}(W|X) = \sum_{V \in \mathcal{M}} p(V|X)\mathcal{L}(V, W),$$

and the *Bayes* decision rule with this cost function is

$$W_{\text{MAP}}(X) = \arg \min_W \sum_{V \in \mathcal{M}} p(V|X)\mathcal{L}(V, W), \quad (1)$$

where the summation set \mathcal{M} includes all non-zero posterior probability word sequences, whereas the minimization goes over all word sequences, including zero-probability word sequences.² In ASR, the standard cost function used in *Bayes* decision rule is the sentence (word sequence, or 0-1) error. Using 0-1 cost, *Bayes* decision rule ends up in the maximization of the class posterior probability, i.e. the maximum-a-posteriori (MAP) rule. With the acoustic feature vector sequences $X = x_1, \dots, x_T$, and word sequences $W = w_1, \dots, w_N$, the MAP for ASR reads

$$W_{\text{MAP}}(X) = \arg \max_W p(W|X).$$

Nevertheless, the choice of a sentence error cost function is inconsistent [5, pp. 4-5] with the evaluation measure used in ASR, the word error rate, which would require a *Levenshtein* cost function to be used in search. In the following, *Bayes* decision rule using *Levenshtein* cost will be called *Levenshtein* rule.

In [8], a strong connection between the MAP and the *Levenshtein* rule was shown. Especially it is shown that if one of the following inequalities is fulfilled, then the MAP and the *Levenshtein* decision rule give the same results:

$$\begin{aligned} 2p(W_{\text{MAP}}|X) + \sum_{\substack{W: \\ \mathcal{L}(W, W_{\text{MAP}}(X))=1}} p(W|X) - \max_{\substack{W: \\ \mathcal{L}(W, W_{\text{MAP}}(X))=1}} p(W|X) &\geq 1 \quad (2) \\ \Leftrightarrow 2p(W_{\text{MAP}}|X) &\geq 1. \quad (3) \end{aligned}$$

As a further specialization, Ineq. (3) (and therefore also Ineq. (2)) follows from the following inequality:

$$\mathcal{R}(W_{\text{MAP}}(X)) = \sum_W p(W|X)\mathcal{L}(W, W_{\text{MAP}}(X)) \leq \frac{1}{2}, \quad (4)$$

since for metric and integer-valued cost \mathcal{L} we have:

$$\begin{aligned} \mathcal{R}(W_{\text{MAP}}(X)) &= \sum_W p(W|X)\mathcal{L}(W, W_{\text{MAP}}(X)) \\ &= \sum_{W \neq W_{\text{MAP}}(X)} p(W|X)\mathcal{L}(W, W_{\text{MAP}}(X)) \\ &\geq \sum_{W \neq W_{\text{MAP}}(X)} p(W|X) \\ &= 1 - p(W_{\text{MAP}}(X)|X). \end{aligned}$$

²In an example in [8, Fig. 1] it was shown that zero probability classes indeed can result from *Bayes* decision rule.

3. Bayes Risk with Word Error Cost

Ideally, all word sequences, including the zero probability word sequences are possible results of *Bayes* decision rule with *Levenshtein* cost (*Levenshtein* rule), cf. [8, Fig. 1]. To reduce complexity, the minimization space in the *Levenshtein* rule, cf. Eq. (1), can be approximated by the summation space \mathcal{M} , i.e. by the non-zero posterior probability word sequences. Let's denote this restricted decision rule as minimum word error (MWE) rule.

In the following we introduce an upper bound to the *Levenshtein* cost function, which provides further insight into the relation between *Bayes* risk and word error rate. A *Levenshtein* alignment between two word sequences can be transformed into a *Hamming* alignment by inserting empty (ϵ) words in first and second word sequence wherever insertion and deletion errors occur, respectively. Also note that the additional insertion of pairs of empty words at the same position would not change the edit-distance between the two aligned sequences. Therefore, a multiple alignment between a given word sequence and a set of word sequences can be arranged in a similar way, where empty words are inserted into each word sequence, such that each aligned word sequence will have equal length (i.e., including the ϵ -words). The first part of Table 1 shows a corresponding example for the *Levenshtein* alignment of word sequences W_i , $i = 1, 2, 3$, to the seed word sequence W_0 using empty words. The second part of Table. 1 shows the *Levenshtein* cost for aligning each pair of these word sequences.

Table 1: Multiple *Levenshtein* alignment of the word sequences $W_1 = a c b$, $W_2 = a d c b$, and $W_3 = a d b$ to the seed word sequence $W_0 = a b c$ represented by a position-wise alignment using inserted empty (ϵ -)words. The table shows the *Levenshtein* cost (edit distance), and the alignment-based cost with the seed sequence W_0 , between all pairs of sequences.

word sequences	<i>Levenshtein</i> cost				multiple alignm. <i>Hamming</i> cost				
	W_0	W_1	W_2	W_3	W_0	W_1	W_2	W_3	
W_0	a	ϵ	b	c	ϵ				
W_1	a	c	b	ϵ	ϵ				
W_2	a	ϵ	d	c	b				
W_3	a	d	b	ϵ	ϵ				

In the following, the representation of multiple *Levenshtein* alignments will be called *multiple alignment* in short. Note that the length of all word sequences within a multiple alignment is made equal by the insertion of empty words. This length will be called the length of a multiple alignment, denoted as $N(W, \mathcal{M})$, with the seed word sequence W and a set of aligned word sequences \mathcal{M} .

To correctly address the individual words of a word sequence $V \in \mathcal{M}$ after a multiple alignment, the notation $V_n^{(W, \mathcal{M})}$ is defined, which refers to the n -th position in word sequence V for a multiple alignment of the set of word sequences \mathcal{M} to word sequence W .

Using this notation, a new word level distance measure can be defined as the *Hamming* distance [3] between word sequences V and W resulting from a multiple alignment of sequences in \mathcal{M} to a seed word sequence U . This distance measure will be referred to as *multiple alignment Hamming cost* in the following:

$$\mathcal{L}^{(U, \mathcal{M})}(V, W) = \sum_{n=1}^{N(U, \mathcal{M})} \ell \left(V_n^{(U, \mathcal{M})}, W_n^{(U, \mathcal{M})} \right), \quad (5)$$

with the local distance $\ell(v, w)$ between two words v, w defined as:

$$\ell(v, w) = 1 - \delta_{v, w} = \ell(w, v). \quad (6)$$

In the third part of Table 1, the multiple alignment *Hamming* cost is shown for the case $U = W_0$ and $\mathcal{M} = \{W_0, W_1, W_2, W_3\}$.

The multiple alignment *Hamming* cost has a clear relation to the *Levenshtein* distance:

$$\begin{aligned} \mathcal{L}(V, W) &= \mathcal{L}^{(W, \mathcal{M})}(V, W) \\ &= \mathcal{L}^{(V, \mathcal{M})}(V, W), \end{aligned} \quad (7)$$

i.e., the multiple alignment *Hamming* cost involving the seed word sequence itself is identical to the corresponding *Levenshtein* cost, as could be seen in the first row and column of the third part of Table 1. The multiple alignment *Hamming* cost can only be larger or equal to the *Levenshtein* distance, due to the optimality of the *Levenshtein* distance:

$$\mathcal{L}^{(U, \mathcal{M})}(V, W) \geq \mathcal{L}(V, W), \quad (8)$$

with the equality (in general not exclusively) obtained both for $W = U$ or $V = U$, cf. Eq. (7). Now define an alignment-dependent *Levenshtein* risk using the multiple alignment *Hamming* cost function, which will be referred to as *Hamming risk* in the following:

$$\begin{aligned} \mathcal{R}^{(U, \mathcal{M})}(W|X) &= \sum_{V \in \mathcal{M}} p(V|X) \cdot \sum_{n=1}^{N(U, \mathcal{M})} \ell \left(V_n^{(U, \mathcal{M})}, W_n^{(U, \mathcal{M})} \right) \\ &= \sum_{n=1}^{N(U, \mathcal{M})} \left(1 - p_n(W_n^{(U, \mathcal{M})}|X) \right). \end{aligned} \quad (9)$$

with the position conditional word posterior probability

$$p_n(w|X) = \sum_{V \in \mathcal{M}: V_n^{(W, \mathcal{M})} = w} p(V|X). \quad (10)$$

Note that the *Hamming* risk depends both on the set of word sequences \mathcal{M} , as well as the seed word sequence U . From Eq. (7) and Ineq. (8) follows:

$$\mathcal{R}^{(W, \mathcal{M})}(W|X) = \mathcal{R}(W|X), \quad (11)$$

$$\mathcal{R}^{(U, \mathcal{M})}(W|X) \geq \mathcal{R}(W|X), \quad (12)$$

i.e. the *Hamming* risk is an upper bound to the *Levenshtein* risk, with equality if the seed word sequence U for the multiple alignment is equal to the word sequence W the risks are evaluated for. Using Eq. (11) and Ineq. (12) we further conclude

$$\begin{aligned} \mathcal{R}(W|X) - \mathcal{R}(U|X) &= \mathcal{R}(W|X) - \mathcal{R}^{(U, \mathcal{M})}(U|X) \\ &\leq \mathcal{R}^{(U, \mathcal{M})}(W|X) - \mathcal{R}^{(U, \mathcal{M})}(U|X), \end{aligned} \quad (13)$$

i.e. a minimization of the *Hamming* risk starting from the seed word sequence U also reduces the *Levenshtein* risk. The *Hamming* risk results from the *Levenshtein* alignment of all word sequences to a specific word sequence U . Note that the computation of the *Levenshtein* risk for word sequence U itself requires the same alignments. The *Hamming* risk can be computed efficiently for all word sequences which were aligned to word sequence U , the complexity is linear in the number of word sequences aligned. Therefore the *Hamming* risk can be used to efficiently check all word sequences for having lower

Levenshtein risk than the original hypothesis U used as alignment seed. Using the MAP word sequence as alignment seed U , this approach was introduced as “ N -best ROVER” [10] (NBR), when applied to the N -best word sequences of a *single* ASR system (instead of multiple systems for the purpose of system combination, as described in [10]). Moreover, NBR can be iterated by consecutively replacing the seed word sequence with the word sequence minimizing the *Hamming* risk, which in the following is denoted as *Hamming* iteration (HIT). Due to Eq. (11), HIT will not increase the *Levenshtein* risk. Although this approach only is locally optimal w.r.t. the *Levenshtein* risk, its complexity (per iteration linear in $|\mathcal{M}|$) is better than minimizing the *Levenshtein* risk directly ($\geq |\mathcal{M}|^2$), assuming that HIT converges well before all possible word sequences are visited. Convergence of HIT means that the optimization of the *Hamming* risk results in the (last) seed word sequence itself, i.e. it can not be improved further.

Note that the HIT rule can result in word sequences which are not contained in the original set of word sequences \mathcal{M} , since in any position a word hypothesis from any of the word sequences contained in \mathcal{M} can be chosen. A hypothesis not contained in \mathcal{M} can still have a *Levenshtein* risk less than its restricted optimum over \mathcal{M} . Also, after application of the HIT rule, the *Hamming* risk of the word sequence resulting from the HIT rule is equal to its *Levenshtein* risk. Neither the HIT rule nor the MWE rule guarantee to find the global optimum.

The relation between the standard MAP rule and *Bayes* decision rule with *Levenshtein* cost presented in Sec. 2 can be translated to the decision rules presented here. If at least one of the Ineqs. (2-4) is fulfilled, then, the MAP, *Levenshtein*, MWE, NBR, and HIT decision rules all coincide. The complete proof of this statement will be presented in a further publication. Nevertheless, the proof is based on the following relation of the word posterior probabilities derived from the multiple alignment (without proof) with the MAP word sequence $W_{\text{MAP}} = w_1^N = w_1 w_2 \dots w_N$ and $u \neq w_n$ in position n :

$$p_n(w_n|X) - p_n(u|X) \quad (14)$$

$$\geq 2p(W_{\text{MAP}}|X) + \sum_{\substack{W: \\ \mathcal{L}(W, W_{\text{MAP}}(X))=1}} p(W|X) - \max_{\substack{W: \\ \mathcal{L}(W, W_{\text{MAP}}(X))=1}} p(W|X) - 1$$

Clearly, if Ineq. (2) (or its special cases, Ineqs.(3) and (4)) is fulfilled, the right side of Ineq. (14) will be greater or equal to zero. Hence, the NBR rule, i.e. the initial iteration of the HIT rule will decide for the words w_n from the MAP word sequence in each position $n = 1, \dots, N$, noting that the optimization of the *Hamming* risk can be done position-wise, cf. Eq. (9), i.e. for each n separately. Sec. 2 also shows that Ineq. (2) leads to equality of MAP and *Levenshtein* rule. Since by definition $W_{\text{MAP}} \in \mathcal{M}$, all decision rules discussed (MAP, *Levenshtein*, MWE, NBR, and HIT) then give identical results.

4. Experiments

Experiments were performed on a number of different speech recognition tasks. The aim was to evaluate the impact of the analytic results given here, and the effect of the word level cost decision rules based on the *Levenshtein* and *Hamming* risk in contrast to the MAP rule. Experiments were performed on the Wall Street Journal 5k (WSJ5k) with setups as in [8] (with a slight correction in the acoustic feature extraction, leading to slightly improved results in this work), the European Parliament Plenary Sessions (EPPS) using the English System 4 (EPPS-EN), and the Spanish System 1 (EPPS-ES), cf. [6]. Task statistics are summarized in Table 2, including the test corpus sizes

by hours, segments, segment length, running words, and the recognizer vocabulary, and baseline word error rate (WER) of each ASR systems used. The experiments with the word-level cost *Bayes* decision rules presented here are based on N -best lists [9] with 10,000 entries extracted from word graphs generated by the baseline ASR systems. The N -best list posteriors are re-normalized to sum to one for each segment, implicitly assuming zero posterior probability for word sequences not contained.

Table 2: Test corpus statistics and baseline performance.

test corpus	audio [h]	number of		PP	vocab. [k]	baseline WER[%]
		segm.	words			
WSJ5k	1.3	740	12137	111	5	3.95
EPPS-ES	6.2	1527	56875	105	61	9.63
EPPS-EN	2.9	644	27386	107	52	9.77

The experimental results are summarized in Tables 3-4. Each corpus is partitioned into sub-corpora, which comprise those segments, for which the following conditions are fulfilled:

- a) *Levenshtein* risk $\leq \frac{1}{2}$, cf. Ineq. (4);
 - b) max. posterior prob. $\geq \frac{1}{2}$, cf. Ineq. (3), but excl. a);
 - c) Ineq. (2) with *Lev.* cost fulfilled, but excl. a) & b);
 - d) Remaining segments, i.e. excl. cases a), b), and c);
- “=” MAP, MWE and HIT decision rules coincide;
“≠” MAP, MWE, and/or HIT decision rules do *not* coincide.

Table 3: Experimental results for the WSJ5k English corpus using N -best list rescoring.

sub-corpus	segm. [%]	words [%]	av. len. [#word]	WER [%]		
				MAP	MWE	HIT
a)	68.2	66.8	16.1	1.43		
b)	8.4	8.3	16.3	6.33		
c)	5.3	5.4	16.7	13.6		
d)	18.1	19.5	17.7	11.0	10.2	10.4
“=”	95.8	95.3	16.3	3.33		
“≠”	4.2	4.7	18.3	16.6	13.2	14.1
all	100.0	100.0	16.4	3.95	3.78	3.83

Table 4: Experimental results for the EPPS Spanish corpus using N -best list rescoring.

sub-corpus	segm. [%]	words [%]	av. len. [#word]	WER [%]		
				MAP	MWE	HIT
a)	26.6	10.8	15.2	3.91		
b)	2.3	1.7	27.5	4.37		
c)	5.5	4.2	28.4	4.99		
d)	65.6	83.3	47.3	10.7	10.6	10.6
“=”	79.4	73.4	34.4	7.94		
“≠”	20.6	26.6	48.1	14.3	13.8	14.0
all	100.0	100.0	37.2	9.63	9.49	9.54

Table 5: Experimental results for the EPPS English corpus using N -best list rescoring.

sub-corpus	segm. [%]	words [%]	av. len. [#word]	WER [%]		
				MAP	MWE	HIT
a)	14.9	5.5	16.1	2.08		
b)	1.4	0.6	16.7	6.00		
c)	3.7	2.7	30.3	3.02		
d)	80.0	91.2	48.5	10.5	10.4	10.4
“=”	73.5	67.5	39.1	8.24		
“≠”	26.6	32.5	39.0	12.9	12.6	12.6
all	100.0	100.0	39.1	9.77	9.67	9.67

Note that both $\{a, \dots, d\}$ and $\{=, \neq\}$ comprise the complete corpus, respectively. The decisions of the MAP, MWE, and HIT rules are proved to be equal if any of the Ineqs. (2)-(4), cf. cases a)-c). Therefore, in these cases only a single WER each is shown in Tables 3-4. Finally, the last row of each of the Tables 3-4 shows the results for the corresponding complete test corpus. The columns of the results tables show the sub-corpus, percentage of segments and running words, average segment length (by number of words), and the word error rates (WER) for the four decision rules MAP, MWE, and HIT as introduced in Secs. 2 and 3.

In addition to cases a)-c), where MAP, MWE, and HIT are proved to coincide (cf. Sec. 2), there still are large subsets of each corpus, for which the equality is not proved, but nevertheless holds, compare cases “d)” and “ \neq ”. In cases, where the MAP and MWE decision are equal, also the word error rates are well below average for all tasks considered here. Therefore, the agreement of MAP and MWE provides confidence information on segment level. The subsets “ \neq ”, where MAP, MWE, and/or HIT differ are comparatively small. Any improvements obtained from word error based decision rules like MWE and HIT originate from these comparably small subsets of the test corpus. This might further explain, why improvements obtained from word error based decision rules usually are comparatively low, as could be observed here, and in a number of other experimental approaches to word error based decision rules, e.g. [4, 7, 9].

Subsets with lower word error rate have a higher percentage of cases a)-c). This could be expected: error rate is linked to the MAP probability via *Bayes* decision rule, and the magnitude of the MAP probability plays an important role for cases a)-c). Although Ineq. (2) covers all three cases a)-c), the bulk of it is already covered by the special case of low risk in Ineq. (4), i.e. case a).

By definition, for cases a)-c) a risk lower than $1/2$, very high MAP probability of segments, or high probability mass for the MAP class and classes with distance one from it are observed. Nevertheless, assuming constant average probability mass *per word position* for the MAP word sequence, the posterior probabilities for the MAP word sequence and similar sequences should decrease with sequence length. Consequently, the average segment lengths for cases a)-c) are lower than the overall average and show a trend, as the word error rates do. This length dependence of the equality conditions for MAP and MWE also supports the choice to cut segments in ASR tasks into smaller segments, as proposed in the experimental approach presented in [2].

The experimental results show that Ineq. (3) already covers a major part of Ineq. (2). Recall that Ineq. (3) follows from Ineq. (4), i.e. it is represented by the union of cases a) and b). Therefore Ineq. (2), following from both Ineqs. (4) and (3) only adds subset c). Using Ineq. (3), in [8] a rough estimate of the ASR word error rate r_1 needed to observe a considerable effect of the MWE over the MAP rule is derived, depending on the average length M of the segments: $r_1 \gtrsim 1 - (1/2)^{\frac{1}{M}}$. Using Ineq. (4), an even simpler, though less tight estimate for the word error rate r can be derived, beyond which a considerable effect can be expected: $r_2 \gtrsim \frac{1}{2M}$. Note that both estimates converge for large M . For WSJ5k, the average segment length is 16.4 words, resulting in $r_1 \gtrsim 4\%$, which is very near to the word error rate observed on WSJ5k. This motivates the low ratio of less than 5% of segments, for which MWE results in different decisions than MAP on WSJ5k, cf. case “ \neq ” in Table 3. On the other hand, for high word error rates and/or long

segment lengths, r would be much higher. Consequently, on the EPPS and GALE corpora, the ratio of segments covered by cases a)-c) is considerably reduced. Also the ratio of segments for which MWE and MAP give different results increases both with average segment length and word error rate.

5. Conclusions

In this work, fundamental conditions were presented, for which *Bayes* decision rule using word error cost and standard sentence (word sequence, 0-1) cost coincides. The analytic results agree with experimental evidence that using a task-specific cost function does not provide considerable improvements on top of the standard, sentence cost based *Bayes* decision rule. As part of the conditions for equality of *Bayes* decision rule with word error and sentence error cost, especially the case of *Bayes* risk lower than one half is interesting, i.e. the case of a conditional average word error per segment lower than a constant. This condition shows the strong relationship between word error rate and the choice of cost function in *Bayes* decision rule. It is consistent with the ASR experiments presented, which show equal decisions using sentence and word error cost, unless the baseline error rates are high. The connection between word error rate and posterior probability can even be traced down to word level. Using the properties of the *Levenshtein* cost, the *Hamming* risk was introduced as an upper bound to the *Levenshtein* risk. The corresponding multiple alignment *Hamming* cost based decision rule is covered by the more general conditions derived in Sec. 2. Therefore, the multiple alignment *Hamming* cost based *Bayes* decision rule also is proved to lead to equal decisions than using *Levenshtein* cost.

6. References

- [1] V. Goel, W. Byrne: “Minimum Bayes Risk Automatic Speech Recognition,” *Computer Speech and Language*, Vol. 14, No. 2, pp. 115–135, 2000.
- [2] V. Goel, S. Kumar, W. Byrne: “Segmental Minimum Bayes-Risk Decoding for Automatic Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 3, pp. 234–249, May 2004.
- [3] R.W. Hamming: “Error Detecting and Error Correcting Codes,” *Bell System Technical Journal*, Vol. 26, No. 2, pp. 147–160, 1950.
- [4] B. Hoffmeister, D. Hillard, S. Hahn, R. Schlüter, M. Ostendorf, H. Ney: “Cross-Site and Intra-Site ASR System Combination: Comparisons on Lattice and 1-Best Methods,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1145–1148, Honolulu, HI, USA, April 2007.
- [5] F. Jelinek: *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1997.
- [6] J. Löff, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney: “The RWTH 2007 TC-STAR Evaluation System for European English and Spanish,” *Proc. InterSpeech*, pp. 2145–2148, Antwerp, Belgium, August 2007.
- [7] L. Mangu, E. Brill, A. Stolcke: “Finding Consensus Among Words: Lattice-Based Word Error Minimization,” *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pp. 495–498, Budapest, Hungary, Sept. 1999.
- [8] R. Schlüter, T. Scharrenbach, V. Steinbiss, H. Ney, “Bayes Risk Minimization using Metric Loss Functions,” in *Proceedings InterSpeech*, pp. 1449–1452, Lisboa, Portugal, September 2005. Note: published version contains errors; corrections are highlighted under <http://www-i6.informatik.rwth-aachen.de/~schluter/BayesMetricLossEUROSPEECH2005.pdf>
- [9] A. Stolcke, Y. König, M. Weintraub: “Explicit Word Error Rate Minimization in N-Best List Rescoring,” *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pp. 163–166, Rhodes, Greece, Sept. 1997.
- [10] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V.R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sönmez, F. Weng, J. Zheng: “The SRI March 2000 Hub-5 Conversational Speech Transcription System,” *Proc. NIST Speech Transcription Workshop*, University of Maryland, MD, May 2000.
- [11] F. Wessel, R. Schlüter, H. Ney: “Explicit Word Error Minimization using Word Hypothesis Posterior Probabilities,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 33–36, Salt Lake City, Utah, May 2001.