DYNAMIC CLASSIFIER COMBINATION IN HYBRID SPEECH RECOGNITION SYSTEMS USING UTTERANCE-LEVEL CONFIDENCE VALUES

Katrin Kirchhoff

AG Angewandte Informatik
Technische Fakultät
Universität Bielefeld, Germany
Postfach 100 131, 33501 Bielefeld, Germany
katrin@techfak-uni-bielefeld.de

Jeff A. Bilmes

International Computer Science Institute
Suite 600
1947 Center Street
94704 Berkeley, USA
bilmes@icsi.berkeley.edu

ABSTRACT

A recent development in the hybrid HMM/ANN speech recognition paradigm is the use of several subword classifiers, each of which provides different information about the speech signal. Although the combining methods have obtained promising results, the strategies so far proposed have been relatively simple. In most cases frame-level subword unit probabilities are combined using an unweighted product or sum rule. In this paper, we argue and empirically demonstrate that the classifier combination approach can benefit from a dynamically weighted combination rule, where the weights are derived from higher-than-frame-level confidence values.

1. INTRODUCTION

In the classical hybrid HMM/ANN speech recognition approach [1] subword unit class probabilities are estimated using artificial neural networks, such as time-delay neural networks (TDNNs) or multilayer perceptrons (MLPs). These probabilities are subsequently converted to scaled likelihoods and are used as state emission probabilities in an HMM-based decoder (see Figure 1). Recently, this approach has been extended by using a combination of different hybrid systems. These systems make use of several neural network classifiers which are based on different input representations and whose outputs are combined before decoding (Figure 2).

The classifier combining approach has repeatedly proven more robust than the standard hybrid approach alone. In [5], four different classifiers for different preprocessing front-ends are combined in the context of robust processing of clean and reverberant speech. Wu et al. [12] combine classifiers that use either RASTA preprocessing or modulation spectrogram [4] features as input representations and in [8], an MLP is used to combine classifiers. In [6], acoustic and articulatory classifiers are successfully combined. All of these works report significant reductions of word error rate, especially under acoustically deteriorated conditions such as noisy and reverberant speech. However, the combination method which has so far primarily been employed is extremely simple: it consists of a product rule which multiplies the individual class-conditional probabilities and normalizes by their priors.

This combination scheme gives equal weight to each of the sub-classifiers at each time frame, regardless of their global or local accuracy. It seems more reasonable to consider a modified combination rule that weights each classifier with respect to the quality of its contribution. Although there are several potential knowledge sources which may be used to derive weighting factors, these possibilities have so far not been reported. One possible knowledge source might be the frame-level accuracy or the frame-level confidence (e.g., entropy) of each classifier. However, frame-level recognition accuracy is not always correlated with final word recognition performance [6]; therefore, frame-level confidence measures do not necessarily guarantee an improved word error rate. As alternative knowledge sources, word-level or utterance-level confidence values could be exploited.

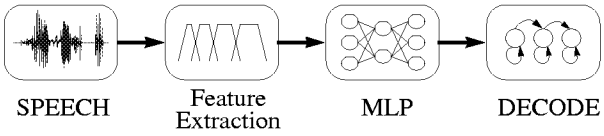

Figure 1: The standard hybrid ANN/HMM approach to ASR.

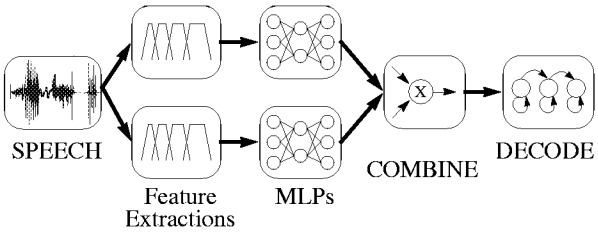

Figure 2: The classifier combination approach considered in this work. The outputs of several ANN posterior probability estimators based on different input representations are combined prior to decoding.

This paper studies the usefulness of various knowledge sources for a dynamically weighted combination scheme. Section 1 describes the speech corpus and the results obtained by our baseline recognizers. Section 2 gives a theoretical analysis of various combination rules and their performance in an initial combination experiment. Section 3 describes "cheating" experiments which lend support to the notion that frame-level information is not necessarily beneficial whereas word-level or utterance-level knowledge can significantly improve recognizer performance. Section 4 describes the development of a confidence tagger and its use in a weighted combination scheme.

2. SPEECH MATERIAL AND BASELINE SYSTEMS

Our experiments were carried out on the OGI Numbers95 corpus [3], which consists of telephone recordings of continuously spoken numbers from a variety of speakers. We use a training set of 3590

System	WER	INS	DEL	SUB
MFCC	6.6	1.0	2.0	3.7
RASTA	7.0	1.6	1.4	3.9

Table 1: Baseline recognition results (in %)

utterances (13873 words), a development test set of 1227 utterances (4757 words) and a test set of 1206 utterances (4673 words). The training and test material has been hand-transcribed at the phone level. Currently, the lowest published word error rate on this test set is 5.1% [13], which was obtained by utterance-level combination (N-best list rescoring) of both a phone recognizer and a syllable recognizer.

Two hybrid ANN/HMM baseline systems were developed for combination: (a) a system using 8 log-RASTA-PLP coefficients, log-energy, and delta coefficients thereof, and (b) a system based on 9 MFCC coefficients plus delta coefficients. In each case, coefficients are computed every 10 ms, using a window of 25 ms. Both systems use a three-layer MLP with 400 units in the hidden layer to estimate 56 output phone probabilities. The input window consists of 9 frames in both cases. Decoding is carried out by a first-best decoder using a back-off bigram language model. Both systems use the same pronunciation lexicon and language model. The language model and acoustic scaling factors were optimized separately for each system.

The baseline recognition results for both systems are shown in Table 1.

3. COMBINATION RULES

The most widely used probability combination rules [7] are the product rule, the sum rule, the min rule and the max rule. Given N classifiers $c_1, ..., c_N$ and K classes $\omega_1, ...\omega_K$, these are defined as follows:

Product rule:

$$P(\omega_k|x_1, ...x_N) = \frac{1}{P(\omega_k)^{N-1}} \prod_{n=1}^{N} P(\omega_k|x_n), \qquad (1)$$

where x_n is the input to the $n'th$ classifier and $P(\omega_k)$ is the a priori probability for class k.

Sum rule:

$$P(\omega_k|x_1, ..., x_N) = \frac{1}{N} \sum_{n=1}^{N} P(\omega_k|x_n) \qquad (2)$$

Min rule:

$$P(\omega_k|x_1, ..., x_N) = \frac{min_n P(\omega_k|x_n)}{\sum_{k=1}^{K} min_n P(\omega_k|x_n)} \qquad (3)$$

Max rule:

$$P(\omega_k|x_1, ..., x_N) = \frac{max_n P(\omega_k|x_n)}{\sum_{k=1}^{K} max_n P(\omega_k|x_n)} \qquad (4)$$

The product and min rule effectively implement an "and" function since the output is large only if both of the inputs are large. The sum rule and the max rule, on the other hand, implement an "or" function since the output is large if either one of the inputs is large.

Rule	WER	INS	DEL	SUB
product	5.3	1.1	1.2	3.0
sum	6.2	0.7	1.9	3.6
min	5.4	1.0	1.4	3.0
max	6.7	0.7	2.0	4.0
avg log	5.4	0.8	1.5	3.1
MLP	6.1	0.9	2.1	3.2

Table 2: Baseline combination results (in %)

We applied all four rules to our baseline recognizers in an initial combination experiment. For comparison, we also investigated an average log-probability combination method which is the N^{th} root of the product rule, and a non-linear combination method, i.e. an MLP which maps the output from both recognizers to the final phone probabilities. The results are shown in Table 2. Obviously, the combination rules which have the effect of an "and" function work best.

4. INTEGRATING WEIGHTING FACTORS

4.1. Cheating Experiments

As we mentioned above, a simple combination rule might benefit from weight factors indicating how strong each classifier's contribution should be. These could be used as exponential coefficients in the product rule:

$$P(\omega_k|x_1, ...x_N) = \frac{1}{P(\omega_k)^{N-1}} \prod_{n=1}^{N} P(\omega_k|x_n)^{\gamma_n}, \qquad (5)$$

where γ_n is the weighting factor for recognizer n. Weighting factors need not be static throughout the decoding process but can change dynamically depending on contextual knowledge.

Intuitively, the more "correct" output should be given a higher weight. In order to determine the usefulness of this concept, we conducted several cheating experiments where "correctness" was defined at several levels. In our first experiment, correctness was defined as the identity of the phone classifier output and the reference label at any given frame. Thus, a "correct" tag was assigned to the frame in case the classifier's highest-valued output coincided with the frame label; otherwise, the "incorrect" tag was assigned. Having obtained this information for both baseline recognizers, equal weights (1.0,1.0) were assigned to the classifier outputs in case they had the same tag. Where tags differed, weighting factors were applied, ranging from (1.0,0.0) (i.e. selection of one system to the exclusion of the other) to (0.5,0.5).

The factors were used in the above weighted product rule (5). Good performance (i.e. frame-level accuracy) of the local ANN classifier in a hybrid recognition system may not necessarily yield a low word error rate. The final recognition performance is to a large extent determined by the interaction of the subword unit distribution estimated by the classifier and the pronunciation models in the recognition lexicon. Therefore, it might be more useful to employ weighting factors indicating which classifier output is more likely to lead to a correct word or utterance. For this reason, we conducted further cheating experiments where classifier correctness was defined as correctness of the word that a given frame belongs to, or correctness of the entire utterance (i.e. 100% word accuracy).

For the word-level and utterance-level experiments, the correctly and incorrectly recognized words/utterances in the individual

Level	WER	INS	DEL	SUB
frame	7.6	2.0	1.5	4.2
word	4.1	0.8	1.3	2.0
utterance	4.2	0.7	1.2	2.3

Table 3: Cheating experiment results in % WER for the frame, word, and utterance knowledge sources, weights = (1.0,0.0)

weight	frame	word	utterance
1.0/0.0	7.6	4.1	4.2
0.9/0.1	5.4	4.7	4.2
0.8/0.2	6.8	4.2	4.4
0.7/0.3	6.5	4.6	4.4
0.6/0.4	6.3	4.8	4.8
0.5/0.5	5.9	5.2	5.1

Table 4: Cheating experiment results in % WER for the frame, word, and utterance knowledge sources using smooth weights.

recognizers' outputs were determined by alignment with the reference transcription. For the word-level experiment, the "correct" or "incorrect" tags were then assigned to each frame in the time segment covered by the word. In the utterance-level experiment, the entire utterance was weighted or de-weighted depending on its tag.

Word error rates for the most restrictive weighting scheme, (1.0/0.0, i.e., selection), are shown in Table 3; results for smoother weighting factors can be found in Table 4. Note that the 0.5/0.5 weighted combination is different than the average log probability method in Table 2; here, weights are not unity only if the tags are different. As we can see, frame-level weighting is not beneficial and in fact has a detrimental effect on product-rule combination. Higher-level information, by contrast, may improve the word error rates by up to 20%.

4.2. Estimating Confidence Values

The goal is how to automatically estimate confidence values indicating recognizer accuracy and how to use these as weights in a combined system. Whereas frame-level confidence estimation and combination can be carried out in a one-stage recognition pass, the use of higher-level confidence values implies delayed combination depending on the temporal context that is used for confidence value estimation. In the case of utterance-level confidence values, classifier combination is carried out after each of the subsystems has finished decoding the utterance; this combination scheme therefore resembles a two-pass decoder.

It is unlikely that confidence values can be estimated with nearly 100% accuracy; however, smooth weights may compensate for deficiencies of the confidence tagger. Since in most of our cheating experiments, utterance-level confidence values gave the best results in combination with smooth weights, we decided to initially concentrate on automatically estimating utterance correctness.

We investigated various decoder features which might serve as relevant predictor variables. As an initial feature, we computed the entropy of the phone posterior probabilities, averaged over the entire utterance:

	average confidence
RASTA	0.938
MFCC	0.944

Table 5: Average utterance confidence values for the two systems.

$$H_{av}(phone) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} p_k^t log(p_k^t) \quad (6)$$

where T is the number of frames in the utterance, K is the number of phones, and p_k^t is the posterior probability of phone k at time t. Additionally, we computed final utterance likelihoods normalized by either the utterance length and/or the average utterance likelihood of each of the recognizers. A linear discriminant analysis showed, however, that the correct and incorrect utterances were not separable solely on the basis of this information.

Since N-best lists were not available from the decoder used for the baseline experiments, we instead used a feature which has previously been proposed for word-level confidence estimation (the feature "astabil" in [9]) and which approximates information from N-best lists: Given a set of decodings which were obtained using different language model and acoustic weight factors ("language model/acoustic jitter"), the "best" decoding is determined and selected as the reference transcription. Each of the alternative hypotheses are aligned against this reference transcription and the number of duplicates is computed for each word and normalized by the number of alternative decodings. A higher score represents higher word confidence. In order to obtain utterance-level confidence values, we averaged the word scores over the utterance.

We computed this utterance confidence feature on both the development and the test set. The hypothesis which obtained the lowest word error rate on the development set was used to define the language model and acoustic weight factors for the "best" decoding of the test set. We found that acoustic jitter has only a minimal effect on the overall confidence scores. For this reason, we decided to limit the number of decodings used to compute utterance confidence to those where just the language model weight was varied, which yielded 30 alternative decodings.

Table 5 shows confidence-based combining results for two general cases. In the first case, we applied non-unity weights depending on which system was more confident (labeled *both* in the table). For example, with the 0.9/0.1 weights, if the RASTA system was more confident about the utterance than the MFCC system, then the RASTA-based phonetic probabilities for all frames of the utterance were give a weight of 0.9. In the second case, non-unity weights were applied only when the "better" system had higher confidence. The MFCC system has slightly higher average utterance confidence scores than the RASTA system (Table 5), and, from the baseline recognition results, it performs slightly better in general. This combination scheme used non-unity weights only for utterances when the MFCC system had higher confidence than the RASTA system. These tests, with a variety of weighting factors, are shown in Table 6, where we obtained a small improvement over our baseline combination method.

Since the use of utterance-level confidence values yielded only a modest improvement, we also investigated word-level confidence values. In order to estimate word correct/incorrect tags from the recognition output, we utilized several features that are commonly mentioned in the literature [11, 2, 10]:

- the duration of the word

	WER	INS	DEL	SUB	weight set
both	5.7	0.8	1.5	3.4	0.9/0.1
both	5.8	0.8	1.6	3.4	0.8/0.2
both	5.8	0.9	1.5	3.4	0.7/0.3
both	5.4	0.7	1.4	3.2	0.6/0.4
MFCC only	5.5	0.9	1.5	3.2	0.9/0.1
MFCC only	5.5	0.9	1.5	3.1	0.8/0.2
MFCC only	5.3	0.9	1.3	3.1	0.7/0.3
MFCC only	5.2	0.9	1.3	3.0	0.6/0.4

Table 6: Combination results using the product rule weighted by utterance-level confidence values.

- language model information, in particular the unigram probability
- the number of states in the word model (indicating not only duration but also the number of pronunciation variants and therefore the confusability)
- the number of times the word was observed in the training set

Most researchers use word-lattice density as a confidence value. Since lattice statistics were not available in our case (we used only a first-best decoder), we instead used the sorted log posterior probabilities of a fixed number of active word hypotheses at the word endpoints, plus the confidence features described above computed at the word level.

This yielded 40 features for each word in each of the recognizers. Information about word correctness was again obtained by forced alignment with the reference transcription. Whereas the beginning and end points of insertions and substitutions were identified simply by comparison with the correct word sequence, deletions were marked based on the comparison of frame-level alignments in order to be able to determine the exact location of the deletion. Both the merged confidence feature vectors for both systems and the correct-incorrect tags were reduplicated for all frames in a given word; the tags assigned to both recognition outputs were then compared frame by frame, and labels were assigned depending on whether both systems were correct, incorrect, or one was better than the other. This data was obtained from the training and the development set. We then trained a "gating" MLP on this data, which had 80 input units, 100 hidden units and 4 output units. The recognition accuracy on the test set was 68%, but the accuracy rates for the cases where one network was better than the other did not exceed 40%. The lowest word error rate obtained using word-level confidence values based on this system was 5.4%.

5. DISCUSSION AND CONCLUSIONS

In this paper, we have demonstrated that accurate higher-than-frame-level confidence values can potentially improve recognition performance using a dynamically weighted frame combination rule. In order to obtain large improvements, however, it is crucial that confidence values are obtained with a high degree of precision.

Several factors contributed to confidence tag inaccuracies. First, we used the standard Numbers95 development test set which has only a limited size. Second, the base-line recognizers were already fairly accurate. Training of the gating network requires many patterns where one sub-system is correct and the other is incorrect. Most of the examples, however, belonged to the "both correct" class since each sub-system gets less than 10% WER. The training data, therefore, was sub-sampled to reduce the chance of the gating network learning only prior probabilities, but this resulted in even less training data. Lack of training material for the crucial cases where one system is better than the other prevented better performance of the network-based confidence tagger. On the other hand, we have demonstrated that simple combination of two similar sub-systems (MFCC and RASTA) leads to a substantial WER reduction; the result is competitive with the best reported result for this database. Furthermore, using language model jitter to produce a dynamic confidence tag and using a one-sided weighting scheme produces a additional small improvement in WER over the baseline.

Future work will use a larger development set and will utilize lattice density or N-best-list measures to produce confidence tags.

6. ACKNOWLEDGMENTS

This work has benefited from discussions with Nelson Morgan. This work has been partially sponsored by ONR URI Grant N00014-92-J-1617 , a DoD IDEA grant, and a grant from the Deutsche Forschungsgemeinschaft under the graduate program "Task-oriented communication" at the University of Bielefeld.

7. REFERENCES

[1] H.A. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer, 1994.

[2] L. Chase. Word and acoustic confidence annotation for large vocabulary speech recognition. *Eurospeech*, pages 815–818, 1997.

[3] R. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. *Eurospeech*, 1:821–824, 1995.

[4] Steven Greenberg and Brian E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, pages 1647–1650. IEEE, April 1997.

[5] B.E.D. Kingsbury and N. Morgan. Recognizing reverberant speech with RASTA-PLP. *ICASSP*, 2:1259–1262, 1997.

[6] K. Kirchhoff. Combining acoustic and articulatory information for speech recognition in noisy and reverberant environments. *ICSLP*, 1998.

[7] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.

[8] Nikki Mirghafori. Multi-band speech recognition: A summary of recent work at ICSI. Technical Report TR97-051, ICSI, Berkeley, CA, USA, 1997.

[9] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. *ICASSP*, pages 875–878, 1997.

[10] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. Neural-network based measures of confidence for word recognition. *ICASSP*, pages 887–890, 1997.

[11] G. Williams and S. Renals. Confidence measures for hybrid HMM/ANN speech recognition. *Eurospeech*, pages 1955–1958, 1997.

[12] S.-L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. *ICASSP*, pages 721–724, 1998.

[13] S.-L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Performance improvements through combining phone- and syllable-scale information in automatic speech recognition. In *ICSLP*, 1998.