

A DYNAMIC CROSS-REFERENCE PRUNING STRATEGY FOR MULTIPLE FEATURE FUSION AT DECODER RUN TIME

Yonghong Yan^{1,2}, Chengyi Zheng¹, Jianping Zhang², Jielin Pan², Jiang Han², Jian Liu²

¹Computer Science & Engineering Department, OGI School of Science & Engineering, Oregon Health & Science University, 20000 NW Walker Rd., Beaverton, OR 97006, USA

²Institute of Acoustics, Chinese Academy of Science, Beijing 100080, P.R. China

{yan,chengyi}@cse.ogi.edu,
{jianping.zhang,jielin.pan,jiang.han,jian.liu}@hcc1.ioa.ac.cn

Abstract

Although multiple cues, such as different signal processing techniques and feature representations, have been used in speech recognition in adverse acoustic environment, how to maximally utilize the benefit of these cues is largely unsolved. In this paper, a novel search strategy is proposed. During parallel decoding of different feature streams, the intermediate outputs are cross-referenced to reduce pruning errors. Experiment results show this method significantly improved recognition performance on a noisy large vocabulary continuous speech task.

1. Introduction

In human speech recognition, various cues (including visual information) are utilized. The more difficult the speech (such as in noisy environments), the more cues are needed [1]. Motivated by how human recognize speech, there has been a strong interest among ASR researchers as to how to combine different features for speech recognition [2,3,4,5,6,7,8,9,10] in recent years. The success of these research approaches is partly due to their efficiency in improving recognition accuracy, partly due to their simplicity and ease of deployment.

Fletcher studied extensively how humans process and recognize speech [11]. This work showed that the phones are processed in independent articulation bands and that these independent estimates are “optimally” merged to achieve the recognition results. Recent research activities on multi-stream or multi-band [3,4,5,6,7,8] also demonstrated the importance of looking at the data from different angles (different signal processing and features) and fusing the information to improve recognition accuracy. However, both Fletcher and the recent activities did not explicitly conclude *how* different information should be fused to form the sound-unit recognition in order to achieve human-like performance.

2. Related work

Traditionally the fusion of different information sources is conducted either by concatenating the feature vectors (such as appending energy to MFCC) or by rescored the merged N-best list or word graph from the output of different streams (post processing). The commonly used ROVER [12] type of approach only fuses information from different sources after the recognitions are completed. The benefit of the concatenated approach (method 1) is that the time dependence of different features can be exploited. Successful examples of this approach include concatenating energy and delta features into the spectral representation (such as MFCC), and multi-stream recognition. The drawback of this approach is that only frame-based features can be incorporated (or only time synchronized feature streams can be incorporated). In multi-stream and multi-band systems, a

set of ANNs is trained for each feature stream and used for probability estimation. The output of these ANNs is combined and input to a HMM decoder as a new set of features. The drawback for this type of approaches is that only frame-based features can be incorporated (that is, only time synchronized feature streams can be incorporated). Segmental based information, such as tones (or pitch patterns), cannot be integrated easily. Mirghafori and Morgan tried to relax the synchrony constraints in their research by using a 2-D HMM [17]. Although in theory 2D HMM and coupled-HMM can be used to address the existing problem, the associated expense is an increased model space (extra states need to be introduced). Although it is straightforward from an implementation point of view, the tremendous increase in the state space dimension makes it impossible for applying to multiple input streams. Attempts were made in [17] but failed to improve accuracy due to significant increase in free parameters that needed to be estimated.

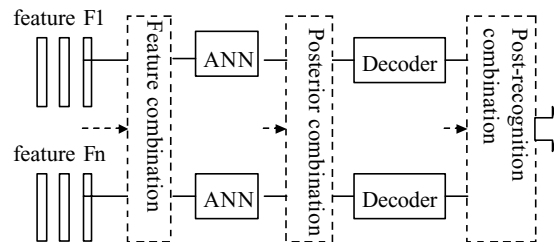


Figure 1: Current fusion strategies in ASR systems

For the post recognition fusion approach the time-dependency between different features is completely ignored during the recognition of each stream. In other words, there is no interaction between different features during the decoding process, i.e. the presence of one feature stream does not impact the course of decoding in other feature streams. The problem with this approach is that possibly complementary information among different feature streams is not fully utilized. The mistakes made early in the decoding stage (or in the first pass) may not be recoverable at the fusion stage (or the second pass processing) since the correct hypothesis may already have been pruned away during decoding for each individual stream. Detailed review can be found in [13]. The other noteworthy post recognition methods are hypotheses combination and confusion networks. Hypothesis combination was introduced in [14]. The word hypotheses obtained from a number of independently trained systems were combined into a word graph. Unlike ROVER, the acoustic score was maintained in each node of the word graph. A language model was used to score the word graph and find the best path as the final hypothesis. This approach tries to explore more paths other than just the first

hypothesis. Confusion Network [15,16] is an approach that aims to minimize the word error rate by post-processing the word graph. It aims to solve the mismatch problem between the current word-based performance criteria and the standard MAP decoding that is sentence-based.

3. The cross reference pruning strategy

In our recent robust speech recognition work [13], rather than focus on discovering a single omnipotent processing or feature representation to solve the acoustic information extraction problem, we focus on how to maximally take advantage of different processing techniques (or feature representations) to make the speech recognition system usable and practical for real applications. The general search process for our proposed work is illustrated in Figure 2. The novel part of our proposed work is the interaction between different feature streams during recognition inside a single decoder framework.

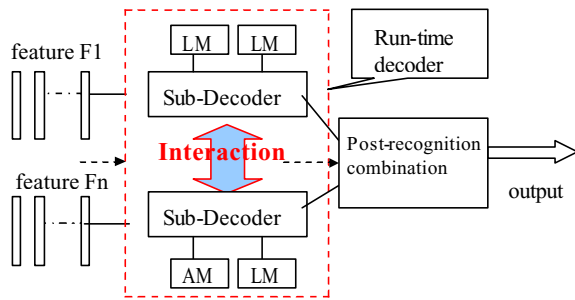


Figure 2: The cross-reference search architecture

Our hypothesis is that, by utilizing the complementary information contained in different feature streams at an earlier stage (such as the decoding stage in recognition), speech recognition accuracy can be greatly improved. The intent is to relax the mathematic constraints in existing approaches and avoid unrecoverable pruning errors at the decoding stage.

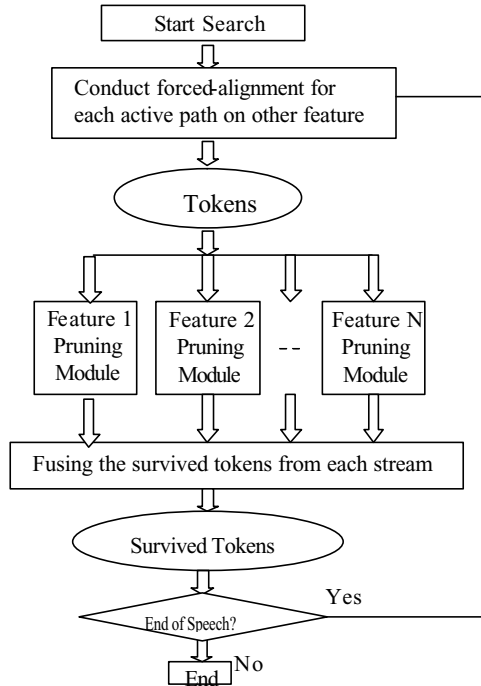


Figure 3: The flow chart of the cross-reference pruning

Mathematically, the proposed approach is different from the previous arts that assume either time-synchrony (concatenating the features to form a single feature stream, such as appending energy to MFCC) or complete independence (running separate recognitions and combining the lattices). The rationale behind this approach is to avoid mistakes caused by pruning based solely on one particular processing/feature representation that fails to capture the characteristics of the underlying phonetic class for a short period of time (such as when particular noise bursts appear or the environment changes).

Based on our previous work on run time information fusion [13], a cross-reference pruning strategy is proposed. The flow chart of this strategy is given in Figure 3. When a pruning decision needs to be made at the fusion point (can be state, phone, word or sentence), if a path is going to be pruned away (fall below the surviving threshold), the same path at other feature streams is referenced. If the corresponding path can survive in different feature streams, then this path will be kept instead of being pruned away. By using features with different time/frequency resolution, the hypothesis is that local minimal can be avoided if particular feature fails to capture the acoustic characteristics of underlying linguistic unit while other features capturing well.

4. Database and system development

The task used to evaluate our approach in this paper is the DARPA SPINE (SPeech In Noisy Environments) task. The second evaluation (SPINE2) was conducted in November 2001 [18, 19]. The test data comprises 128 speaker-environment pairs with 8 different noise environments. The test data has unseen speakers and noise types from the training data, so there will be unavoidable speaker and environment mismatch between the training and test data. The test data contains 3.2 hours speech with 24,015 reference words.

Our SPINE2 system used our Large Vocabulary Continuous Speech Recognition (LVCSR) system [23,13]. It uses decision tree based context clustering, and supports within word and cross word context-dependent phonemes (triphones). The decoder uses a two pass search strategy: the first pass generates a word graph using a simpler acoustic model (within word triphones) and language model (bigram); the second pass re-scores the word graph using a more detailed acoustic model (cross word triphone) and language model (trigram). The system used the same development method as outlined in [13]. The new additions are embedded class language model and speaker adaptation. The best official evaluation results of SPINE2 for using common language model and unrestricted language model are 38.1% and 28.0% WER respectively [22].

5. Experiments & results

Three features, MFCC, TRAPS [20] and TLDA [21], were used in our experiments. Three sets of acoustic models were also trained for each feature based on the same training data.

5.1 Fusion on reducing WER

A word level Constraint Fusion [13] was used as the platform for testing the proposed cross-reference pruning strategy. In our implementation, once a feature is selected as the main feature, the rest two features will be served as consultants (supporting feature) to the selected main feature. During word level token pruning, when the decision is to keep a token (survive) under

normal pruning strategy (in our implementation, the difference between the likelihood of the max token at current time frame and the likelihood of this particular token is compared to a preset threshold), no consultation is made to the rest two features. However if the decision is to prune away the path, a consultation is made via cross-referencing the competitiveness of the same path in the search spaces of the supporting features under the normal pruning strategy. If the path could survive in the supporting feature spaces, then the path will be kept in the main feature search space. The results of a number of experiments are summarized in Table 1.

	MFCC	TLDA	TRAPS
Baseline	26.7%	27.6%	28.9%
Run-Time Fusion	24.3%	25.3%	26.5%

Table 1: WER reduction by using fusion approach

In table 1, the columns MFCC, TLDA and TRAPS denotes the experiment that the corresponding features are selected as the main feature. Baseline row denotes systems that do not apply the proposed cross-referencing pruning strategy and the fusion row denotes systems that applied the proposed strategy.

	MFCC	TLDA	TRAPS	ROVER
Baseline	27.7%	28.6%	29.9%	27.3%
Run-Time Fusion	25.8%	26.7%	27.7%	25.5%

Table 2: Further WER reduction by applying ROVER on the 3 feature systems

Since our run-time fusion approach was performed at different stage compared with pre- and post-recognition fusions. These approaches can be combined in a sequential way. For example, we further conducted ROVER on the outputs from those 3 features systems in Table 1 on both the baseline and fusion conditions. As shown in Table 2 the proposed strategy consistently outperformed baseline results. It also shows that the benefit of our approach can be successfully combined with post-recognition fusion such as ROVER.

5.2 Fusion in dynamic beam adjustment

In large vocabulary speech recognition, the potential search space is prohibitive for a full search. Beam search is a must to limit the search space by pruning away those less likely tokens. In the synchronous search framework, beam search means at every time frame, only the most promising tokens are retained. The recognizer will produce a tremendous amount of token when decoding unintelligible speech segments. For tasks containing significant amount of noise or spontaneous speech, this condition is often occurs. To accelerate decoding under such condition without any damage to the recognition accuracy, we proposed a dynamic beam adjustment in our run-time fusion scheme. In this approach, all features are involved into the pruning decisions. Each feature has its own set of beam pruning values, some of those beam values are dependent on each other such as total allowed token numbers, but most of them are independent. The tokens are first judged by each feature's pruning module, their results are fused to make the final pruning decision. At each time frame, we gather some statistical information on the beam pruning for each feature. For example, at time frame t , word-end beam width of feature f_i will keep α_i percentage of tokens active. The word-end beam width for these features at time

frame $t+1$ will be adjusted according to the following formulas:

$$\varphi = \text{Min}_i(\alpha_i) * \beta; \quad (1)$$

$$\text{WordEndBeam}_i = \text{WordEndBeam}_i * (1 - \frac{\alpha_i - \varphi}{\varphi}); \quad (2)$$

When the survived token percentage of feature f_i over/less than φ , we reduce/increase its pruning widths by certain factor according to equation 2. This approach is based on the rational that unintelligible speech segments are often occurred in some continuous frames. Rather than let lots of unpromising tokens go through each pruning stage till the final histogram pruning, we tried to reduce the active tokens in the earlier stage of pruning. Experiments show this approach significantly speeds up the pruning effort without any loss of recognition accuracy. Figure 4 compares the active token numbers of dynamic beam adjustment with conventional single feature pruning. The curve of fusion approach is relatively flat and the token numbers are in the range of 10k to 30k. All other 3 features have some peaks over 40k tokens.

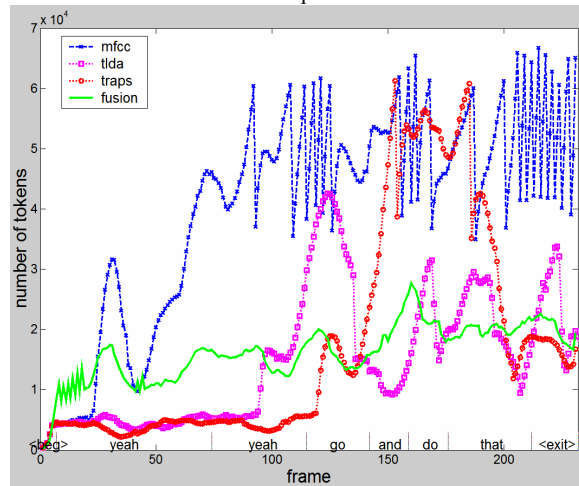


Figure 4: The comparison of active token numbers along the time frame during decoding one sentence

5.3 Fusion in improving word graph quality

The Graph Word Error Rate (GWER) [24] decides the performance upper bound a second pass decoding or rescoring can reach. A word graph with lower GWER can improve the performance of other post-recognition approaches [14,15,16]. Compared to tasks with similar baseline WERs, the SPINE task has a much higher GWER; it is one unique character that distinguished itself with other tasks. As reported in [13], we were able to reduce GWER from 24.0% to 20.9% by using our constraint fusion method, but it is still unsatisfied. We further reduced the GWER by increasing the search beam width. Although the GWER was reduced, it soon reached the limitation of our computation power especially the memory.

GER	WGD	x RT (Real Time)
20.9%	95.3	3.5
19.7%	231.5	6
16.6%	255.2	7.5
14.1%	290.2	10

Table 3: The effect of beam width pruning on word graph size and its accuracy; WGD is a measurement of word graph size

It is important to improve word graph quality without incurring significant computation cost increase. Our cross-reference pruning coupled with the proposed dynamic beam adjustment strategy, at a similar Word Graph Density (WGD), significantly reduced the GWER by about 43% as shown in Table 4. Please note that our new adopted class language model has significantly improved the word graph quality compared to the common language model [13].

	GWER	WGD
Baseline	16.6%	200.2
Fused Token Pruning	9.4%	200.6

Table 4: Graph Word Error Rate comparison

A 2nd pass cross-word decoding was performed on the above word graphs using a same set of acoustic and language models. The improved word graph (Table 4) produced by the 1st pass decoding gave a 9.7% WER reduction in the 2nd pass decoding (Table 5).

	GWER	WER
Baseline	16.6%	26.9%
Fused Token Pruning	9.4%	24.3%

Table 5: The effect of GWER on the WER of 2nd pass decoding

6. Conclusions

In this paper we proposed a pruning strategy for multiple feature information fusion at decoder run time. Experiments showed that it outperforms baseline system consistently across different feature representations. The results further show that it does not cancel the gain by post recognition processing. This best result 24.3% compared favorably to the best official evaluation result 28.0% [22]. Our approach is also flexible enough to perform higher accuracy beam pruning, resulting in faster and more accurate decoding. A significant improvement on the word graph quality is obtained and further benefit is observed on a 2nd pass decoding. At current stage of our research for dynamic information fusion, only search is being investigated. In order to obtain the maximal benefit of using multiple features, a jointly training scheme is also needed. We will investigate the possible extension of the EM algorithm so that HMMs for different features can be trained simultaneously and jointly to ensure global optimization.

7. Acknowledgement

The authors thank Prof. Hynek Hermansky and his students for providing the three features (TLDA and TRAPS) that used in this work. This work is partially supported by ETIC of Oregon University System and Institute of Acoustics, Chinese Academy of Science.

8. References

- [1] P.D. Denes, E.N. Pinson. *The Speech Chain: The physics and biology of spoken language*, W.H. Freeman and Company, New York, 1993.
- [2] Y. Yan, E. Barnard. An Approach to Automatic Language Identification based on Language-dependent Phone Recognition. Proc. of ICASSP, 1995.
- [3] H. Christensen, B. Lindberg, et al. Employing Heterogeneous Information in a Multi-stream Framework. Proc. of ICASSP, 2000.
- [4] A.K. Hallberstadt and J. R. Glass. Heterogeneous Measurements and Multiple Classifiers for Speech Recognition. Proc. of ICSLP, 1998.
- [5] K. Kirchhoff and J. A. Bilmes. Dynamic Classifier Combination in Hybrid Speech Recognition Systems Using Utterance-Level Confidence Values, Proc. of ICASSP, 1999.
- [6] A. Janin, D. Ellis and N. Morgan. Multi-Stream Speech Recognition: Read for Prime Time? Proc. of Eurospeech, 1999.
- [7] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas. On combining classifiers. *IEEE Trans on Pattern analysis and Machine Intelligence*, 20:226-239, 1998.
- [8] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. Proc. ICSLP, 1998.
- [9] R.A. Jacobs, M. Jordan, S. Nowland and G. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3:79-87, 1994.
- [10] D. Ellis and J. Bilmes, Using mutual information to design feature combinations, Proc. of ICSLP, 2000.
- [11] J. B. Allen, How do humans process and recognize speech? *IEEE Trans on Speech and Audio Processing*, pp. 567-577, Vol 2, No. 4, Oct 1994.
- [12] J.G. Fiscus, A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). Proc. of IEEE Workshop on ASRU, 1997.
- [13] C. Zheng, Y. Yan, Run time Information fusion in speech recognition, Proc. of ICSLP, 2002.
- [14] R. Singh, M.L. Seltzer, B. Raj and R.M. Stern. Speech in Noisy Environment: Robust Automatic Segmentation, Feature Extraction and Hypothesis Combination, Proc. of ICASSP, 2001.
- [15] L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words: Lattice-based word error minimization. Proc. of Eurospeech, 1999.
- [16] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks, *Computer Speech and Language*, 14, 373-400, 2000.
- [17] N. Mirghafori, N. Morgan, Combining Connectionist Multi-Band and Full-Band Probability Streams for Speech Recognition of Natural Numbers, Proc. of ICSLP, 1998.
- [18] Naval Research Laboratory (NRL). SPEECH In Noise Environment. <http://elazar.itd.nrl.navy.mil/spine>. 2001.
- [19] A. Schmidt-Nielsen, T. H. Crystal & E. Marsh, Speech in Noisy Environments (SPINE) Adds New Dimension to Speech Recognition R&D, Proc. of HLT, 2002.
- [20] H. Hermansky, S. Sharma. Temporal patterns (TRAPS) in ASR of noisy speech. Proc. of ICASSP, 1999.
- [21] S. Kajarekar and H. Hermansky. A Study of Two Dimensional Linear Discriminants for ASR. Proc. of ICASSP, 2001.
- [22] V.R.R. Gadde *et al*, Building an ASR for Noisy Environment: SRI's 2001 SPINE Evaluation System. Proc of ICSLP, 2002.
- [23] Y. Yan, X. Wu, J. Schalkwyk and R. Cole. Development of CSLU LVCSR: the 1997 DARPA Hub4 evaluation system, Proc. of DARPA 1998 BNTUW workshop, 1998.
- [24] S. Ortmanns, H. Ney, X. Aubert. A Word Graph Algorithms for Large Vocabulary Continuous Speech Recognition. *Computer Speech and Language*, 11, 43-72, 1997.