INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A WFST Framework for Single-Pass Multi-Stream Decoding

Sirui Xu and Eric Fosler-Lussier

Department of Computer Science and Engineering,
The Ohio State University,
Columbus, OH 43220
xu.359@osu.edu, fosler-lussier.1@osu.edu

Abstract

Combining disparate automatic speech recognition systems has long been an important strategy to improve recognition accuracy. Typically, each system requires a separate decoder; final results are derived by combining hypotheses from multiple lattices, necessitating multiple passes of decoding. We propose a novel Weighted Finite State Transducer (WFST) framework for integrating disparate systems. Our framework is different from the current popular system combination techniques in that the combination is done in one-pass decoding and allows the flexibility to combine systems at different levels of the decoding pipeline. Initial experiments with the framework achieved comparable performance as MBR-based combination which is reported to outperform ROVER and Confusion Network Combination (CNC). In this paper, we describe our methodology and present pilot study results for combining systems that use different sets of acoustic models, 1) gender-dependent GMM models, 2) MFCC and PLP features with GMM models, 3) MFCC, PLP and Filter Bank features with DNN models, and 4) SNR-specific DNN acoustic models. For each experiment, we also compared the computation time of the combined systems with their corresponding baseline systems. Our results show encouraging benefits of using the proposed framework to improve recognition performance while reducing computation time.

Index Terms: Acoustic Model Combination, WFST, Semirings

1. Introduction

Research on automatic speech recognition (ASR) has witnessed substantial development over the past few decades. For many years, the use of Hidden Markov Models (HMMs) in combination with Gaussian Mixture Models (GMMs) was the dominant method for acoustic modeling, which achieved considerable improvement in recognition accuracy. Recently, the introduction of Deep Neural Networks (DNNS) has further improved the recognition performance.

In a typical ASR system, multiple stages are often employed. Each stage uses the output of the previous stage to continue processing, thus forming a sequential ASR pipeline. At the end of the pipeline is the decoding process. During this stage, n-best results are kept and the pruning threshold is carefully chosen. Models that are used can be represented as graphs, which makes weighted finite-state transducers (WFST) an efficient representation of these models. WFST-based decoders [1] admit to a convenient compositional structure, where information about the acoustic, context, pronunciation, and language models can be integrated. Operations such as composition, determinization and minimization enable WFST to manipulate the intermediate graphs for optimal results.

At the heart of the WFST is the semiring that governs how the probabilistic scores are propagated through combined network. The traditional WFST decoder implementing the Viterbi algorithm utilizes the tropical semiring, which operates in negative log space to find the best decoding path that minimizes the cost (or, conversely, finds the most probable utterance). In recent years, different types of semirings have been proposed for various purposes. For example, the lexicographic semiring by Shafran et al. [2] is used for determinizing tagged word lattices. Also, van Dalen et al. [3] report the use of expectation semiring to efficiently extract features.

In ASR research, system combination has been a focused area of investigation. Different approaches have been proposed to combine systems to utilize the advantages of each system for better performance. Representative of these are ROVER [4], Confusion Network Combination (CNC) [5], and Multi-Stream Combination [6, 7, 8, 9]. Recently, the Minimum Bayes Risk (MBR) combination method proposed by [10] is reported to outperform the more traditional ROVER and CNC. However, most of these system combination techniques perform multi-pass decoding, which makes the decoding process more complex and time consuming.

In this paper, we propose a WFST framework that extends the traditional semiring into vector semirings, which permit combination of multiple acoustic models in the decoding phase to achieve better single-pass recognition performance. In Section 2, we introduce the vector semiring and decoding graph generation process. In Section 3, we describe the pilot experiments conducted. We combined GMM-based acoustic models based on two genders, MFCC and PLP features with GMM models, as well as MFCC, PLP and Filter Bank features with DNN models. We also explored the possibility of combining models for different noise environments. Our results show performance comparable to the MBR-based framework [10], despite combining on the uttterance, rather than word, level. For each experiment, we also assessed the ability of the combined systems in reducing computation time. In the final section, we discuss the results and future work.

2. Methods

One popular way to combine different systems is the ROVER technique [4]. This method generates a Word Transition Network (WTN) from a set of hypotheses of different systems and then uses majority voting to produce a single recognition hypothesis. Other approaches such as Confustion Network Combination (CNC) [11] and Lattice Combination [12] also rely on the generation of word-level lattices or networks which are products of the decoders. These approaches have achieved improvement in ASR and keyword search areas. In addition,

http://dx.doi.org/10.21437/Interspeech.2016-1307

	Log vector semiring	Tropical vector semiring
Set:	$\mathbb{R}^n \cup \pm\infty$	$\mathbb{R}^n \cup \pm\infty$
$(a_1, \ldots, a_n) \oplus (b_1, \ldots b_n)$	$(a_1 \oplus_{log} b_1, \ldots, a_n \oplus_{log} b_n)$	$\begin{cases} (a_1, \ldots, a_n), \text{if} \min(a_1, \ldots, a_n) \leq \min(b_1, \ldots, b_n) \\ (b_1, \ldots, b_n), \text{if} \min(a_1, \ldots, a_n) > \min(b_1, \ldots, b_n) \end{cases}$
$(a_1, \ldots, a_n) \otimes (b_1, \ldots b_n)$	$(a_1 + b_1, \ldots, a_n + b_n)$	$(a_1 + b_1, \ldots, a_n + b_n)$
$\mathbf{0}$	$(+\infty, \ldots, +\infty)$	$(+\infty, \ldots, +\infty)$
$\mathbf{1}$	$(0, \ldots, 0)$	$(0, \ldots, 0)$

Figure 1: Definition of vector semirings for both log and tropical domains.

multi-stream techniques [6, 7, 8, 9] have been used to combine acoustic scores from different models at the frame level. Different techniques can be employed to combine acoustic scores, such as taking maximum or average or using MLP [13]. Joint decoding, another approach of combining systems, tries to incorporate different forms of an acoustic model into the decoding process. For example, recent work by Wang et al.[14] tried out joint decoding in keyword search tasks. In their work, a Tandem and a Hybrid acoustic model, which share the same HMM structure and decision tree, are combined together.

Currently, WFSTs are the dominant structure for ASR decoding, but there is limited research on combining different acoustic models under WFSTs. A standard WFST provides a unified framework for speech recognition systems to represent different knowledge sources, i.e. the acoustic model, context model, pronunciation model and language model, which are generated as separate WFSTs: H, C, L, G, and then composed and optimized with WFST operations to obtain the final graph for the decoding phase. Each internal arc in a WFST carries a single weight corresponding to the negative log probability of the arc, and a single label (for H, G) or label pair (for C, L) to build the probabilistic word sequence.

Because standard WFSTs only carry a single weight and a single input label, it is not obvious how to apply combination of multiple acoustic models. When considering multi-stream decoding, one might choose to combine acoustic model information at the frame level (as in [14, 15], *inter alia*), which does not require changing the WFST decoding structure. A system that allows for model combination to happen at a longer timeframe under WFST will need to modify the WFST structure. For example, [16] employed a Multi-Tape FST to combine features generated at variable rates.

So far, there is not an effective framework to combine different acoustic models that have distinct decision tree structures at different levels of the speech recognition chain. To do this, it requires a vector of partial hypothesis scores and labels to be carried along the WFST decoding graph and pass through the decoding process. In this paper, we propose a framework to achieve such kind of combinations.

Our framework integrates different trained acoustic models into the decoding process by extending the one-label WFST lattice to include multiple labels and weights on the arcs of WFST. Each of the labels comes from a separate acoustic model, and represents a tied tri-phone state (senone). While decoding, all labels will be kept on the WFST till the end of decoding. Then the labels corresponding to the best path will be chosen to form the final hypothesis. We refer to this extended WFST as parallel-labeled WFST. Since there are multiple labels and weights on the arc, in order to do WFST computations, we need to extend the traditional single-weight (scalar) semiring to a vector semiring. We build on Shafran et al.'s idea [2] of the lexicographic semiring which carries multiple weights, but our weights correspond to scores from multiple acoustic models.

2.1. Vector semirings

In contrast to the scalar semiring defined over $\mathbb{R}$, a vector semiring is defined over $\mathbb{R}^n$. As an example, we extend the scalar log semiring and scalar tropical semiring as shown in Figure 1.

In the definition of the tropical vector semiring, the $\oplus$ operation is extended to a vector space to select the minimum value between vectors. It is important to note that there are different ways to define the $\oplus$ operation: one could define $\oplus$ to pick the vector with the minimum norm, or to pick the point-wise minimum of each element in the vectors; the space of possible $\oplus$ operations is a point of future exploration. The particular choice of $\oplus$ we use here is discussed in the next section.

2.2. WFST graph generation and decoding

To extend the WFST, we first train multiple acoustic models with different data sources or different forms of the same data. Because the models are trained separately, we will have different decision trees and therefore different outputs for a single acoustic unit. For example, in our experiments with triphones, each acoustic model would generate a different senone label for a single frame, and we combined all the senone labels together to form an integrated H transducer. Each arc in the H transducer now carries multiple senone labels in parallel and their respective weights. We also extend the C, L, and G transducers to carry multiple labels and weights. Finally, we compose these parallelized H, C, L, G transducers and determinize and minimize to obtain the final WFST.

The combination can be achieved at different levels of the decoding graph $H \circ C \circ L \circ G$. In our pilot experiments, we combine at the utterance level, which is theoretically equivalent to choosing the maximum likelihood model. During the decoding process, the scores of the multiple acoustic models are accumulated separately, and at the end of one whole utterance, the one with the best score is picked for calculating the hypothesis of the utterance. The particular choice of the tropical vector semiring $\oplus$ (Section 2.1) allows us to implement selection of the best score.

Key to this process is a new operation, FOLD, which can be thought of as a process to take an FST W_1 defined on vector semiring s_1, and return FST W_2 defined on scalar semiring s_2:

$$FOLD(W_1, s_1, s_2) \rightarrow W_2.$$

In essence, the FOLD function maps a graph with multiple weights to another graph with a single weight. Therefore, the combination of two acoustic models at the utterance level can be formally represented as

$$FOLD(< H_1, H_2 > \circ C \circ L \circ G, s_1, s_2),$$

where H1 and H2 are the two acoustic models to be combined,

s_1 is a two-weight tropical semiring and s_2 is a scalar tropical semiring. The resulting graph will be a WFST with single weight.

There is more than one way to define how the FOLD function maps the weights. Choosing the hypothesis with the highest likelihood corresponds to the max operation. However, one can consider more complex operations that implement other combination methods (e.g., Minimum Bayes Risk).

We can also imagine that the combination can be done at different levels of the decoding chain, e.g. the word level:

$$FOLD(< H_1, H_2 > \circ C \circ L, s_1, s_2) \circ G,$$

While theoretically our pilot experiments choose the maximum probability hypothesis over all models, in practice there may be interactions with pruning: parallel decoding may change the likelihood of pruning out the correct hypothesis. For example, when multiple models are used, even if the score of one of the models is lower than the pruning threshold, the correct hypothesis can still be kept if the scores of the other models are higher than the pruning threshold. It is unclear to what degree this is actually an issue.

3. Data Set Description

To evaluate our framework for acoustic model combination, four sets of experiments were conducted on Switchboard [17] and CHiME 2 [18] datasets with the Kaldi toolkit [19].

For Switchboard dataset, the JHU WS96 split is used to divide the data. There are in total 75888 utterances in the training set and 409 utterances in the evaluation set. After the removal of duplications and utterances that are too short, 50868 utterances are selected from the training set to train the models.

The CHiME 2 dataset is constructed based on the Wall Street Journal 5000-word vocabulary speech corpus. The training set contains 7138 utterances recorded from 83 different speakers; a randomly selected SNR within the range of -6 to 9 dB is applied to each of the utterances. The development set includes 409 noisy utterances from 10 speakers, while the evaluation set consists of 330 noisy utterances from 8 other speakers.

4. Experimental Setup

We conducted four pilot experiments to examine our proposed framework in combining different acoustic models. For each experiment, we also calculated the ratio of computation time of each combined system to their corresponding baseline system to see if our proposed framework is able to save time.

4.1. Gender-dependent Switchboard

Our first pilot experiment uses two acoustic models trained separately for males (26354 utterances) and females (24514 utterances). Two standard HMM/GMM models were trained in Kaldi for the two genders and then combined using the multi-stream decoder. The baseline system for comparison was a speaker independent system with standard HMM/GMM setup. MFCC delta + double-delta features were used for all systems and no speaker adaptation or discriminative training was used.

4.2. MFCC and PLP features with GMM models

A second experiment tests the ability to use models based on different acoustic features. MFCC and PLP features are the two commonly used features in the field. PLP features are more robust when there is an environmental mismatch between the training and testing data. MFCC features, on the other hand, can perform slightly better than PLP when training and testing data are both taken from a clean environment.

During our experiments, we trained separate GMM acoustic models for MFCC and PLP features using the same full Switchboard training set. These two models were then combined and the results were compared with the model trained only on MFCC or PLP respectively.

4.3. MFCC, PLP and filter bank features with DNN models

The third set of experiments repeats (and expands) the previous experiment, but with more state-of-the-art DNN acoustic models. We combined three types of features: MFCC, PLP and filter bank features. MFCC and PLP features were transformed with LDA+FMLLR transformations, which were computed from previous GMM-HMM system. For each of the three features, we trained separate DNNs and used our framework to combine the three DNN systems.

The DNN training consisted of two steps. Firstly, we trained the DNNs with the cross-entropy criterion. The features were transformed to have zero mean and unit variance before they were used as the input of the DNNs. Then the state-level minimum Bayes risk (sMBR) criterion was used to train the final set of DNNs. The previous cross-entropy DNNs were used as the starting point for the training of sMBR DNNs, and the training of the DNNs was started from using a uni-gram language model to generate the lattices.

4.4. Combining DNN models for different noise levels using CHiME 2 dataset

The baseline system we used for this set of experiments is a 7-layer DNN model trained with the whole CHiME 2 training dataset and the sMBR criterion. Our constrast models were developed by training 6 separate DNN models for data corresponding to one SNR level. MFCC features were used to obtain GMM-HMM models, based on which 7-layer DNN models were then trained with filter-bank features. As above, the sMBR criterion was used during DNN training.

5. Pilot Results

In this section, we present the preliminary results obtained from our experiments. Table 1 shows the baseline model WER, multi-stream combined system WER, best likelihood for each model and MBR combination WER. We include best likelihood WER to show the best possible result that can be achieved from multiple systems without the use of combination techniques; in the theoretical case of no pruning during decoding, the best likelihood and multi-stream systems should match given the current definition of the $\oplus$ combination operation used in this paper (max probability). The WER difference between the two systems reflects the interaction of the vector semiring with the pruning in the decode process. In the columns of multi-stream WER and best likelihood WER, we also report the ratio of computation time to a single independent baseline model. We also use the MBR combination technique as comparison to validate our framework, since the MBR combination technique is reported to show improved performance than traditional ROVER and CNC [10]. The MBR technique is able to find hypothesized word sequences that minimize the Bayes' risk of producing the wrong word by combining the lattices generated by different

Experiments	Baseline WER	Sub-systems Combined	Multi-Stream WER (Time rel. to baseline) *proposed method*	Best Likelihood WER (Time rel. to baseline)	MBR Combination
Gender Dependent SWB GMM	40.0%	Male+Female (MFCC)	39.6% (1.0x)	41.2% (2.0x)	39.8%
MFCC/PLP SWB GMM	MFCC 40.0%	MFCC+PLP	39.9% (1.1x)	39.8% (2.0x)	39.8%
	PLP 40.5%				
MFCC/PLP/fBank SWB DNN	MFCC 26.4%	MFCC+PLP	26.3% (1.1x)	26.2% (2.0x)	26.1%
	PLP 27.5%	MFCC+fBank	26.2% (1.1x)	26.1% (2.0x)	26.1%
	fBank 26.9%	MFCC+PLP +fBank%	26.1% (1.1x)	26.1% (3.0x)	25.9%
SNR dependent CHiME2 DNN	21.1%	6 SNR Levels	20.6% (2.4x)	20.4% (6.1x)	20.5%

Table 1: Word error rates (WER) and relative decode times for a single (baseline) model and multi-stream one pass decoder combining multiple models. Comparisons are provided with systems that run multiple decoders and choose the output with either Best Likelihood at the utterance level and MBR-based lattice combination.

systems. Because the MBR system allows combination at the word level, we expect that this would be a top-line estimate of performance on these tasks.

As is seen in Table 1, across the set of experiments, there is typically some improvement by combining acoustic models; the improvement is not very dramatic but is relatively consistent, which helps to validate our proposed framework, as the error rates for the multi-stream decoder are almost always within $\pm 0.2\%$ of the result from multiple decoders. The best result comes from the CHiME-2 experiment, with about 0.5% improvement on the test set using SNR-dependent models.

We note that the accuracy improvement of the combined systems relies on whether the different systems are able to provide complementary information. As the results show, the gender dependent combined system showed higher improvement than the systems that combined acoustic models from different features (i.e. MFCC/PLP GMM and MFCC/PLP/fBank DNN). This may be because the features we used in the experiments carry similar information, which limits the space for improvement in the combined systems. To some degree, we expected these results because taking the result from the highest scoring model is a relatively straightforward combination method, especially compared to typical lattice combination techniques. However, the experiment results showed that the performance of our proposed framework was also comparable with the MBR combination technique, which is encouraging given the simplicity of the combination rule used here.

The most critical advantage of the approach is the time saving observed using single-pass models over comparable multiple decodes. As Table 1 shows, the combined systems all demonstrate much shorter decode times compared with the baseline models.[1] For the combination of two models (two genders and two features), the computation time of the combined systems is very close to the traditional single acoustic systems. For the combination of three models (MFCC, PLP, Filter Bank), the computation time of the combined system showed a considerable reduction compared with the single baseline model. For the six-model combination on the CHiME 2 dataset, the reduction of computation time was significant: more than a factor of two faster over decoding separately. Parallelizing the decoding processes through the multi-stream WFST decoder can save considerable computation time over multiple decodes.

Additionally, we also calculated the size of WFST used for decoding. For most of the experiments, the WFSTs of the combined systems have almost the same size as their corresponding single model baseline system. Even for the 6-model combined system on CHiME2 dataset, the increase of the number of nodes and arcs is about 10%. In the experiments, all the combined decoding graphs were easily fit into memory on standard linux servers, and we did not find dramatic increase in memory usage.

6. Conclusions

In this paper, we describe an extension of the WFST framework, using vector semirings, to handle multi-stream decoding and present four sets of experiments examining the effectiveness of combining models within this framework. Our proposed framework has two major advantages: 1) it allows the flexibility to combine multiple systems at different levels of the decoding pipeline (e.g. frame, subword, word and utterance level); 2) it allows one-pass decoding, which is simpler than traditional lattice combination methods. As shown in our pilot experiments, the framework achieves improved accuracy while efficiently reducing computation time. The proposed system is relatively memory-efficient as well, even when decoding six streams simultaneously, because of the shared decoding structure between the streams.

We plan in future work to extend this approach in two directions: first, we plan to examine a richer set of combination functions for the FOLD operation (as well as other design choices, such as the definition of the $\oplus$ operator for the vector semiring). The pilot experiments that were conducted in this paper used the most straightforward combination method (choose the maximum likelihood utterance), and we plan to explore different alternatives that allow for cross-fertilization of hypotheses from different streams.

A second thread of work concerns experiments on the selection of the optimized result at the word or sub-word level rather than at the utterance level (as used in this paper). This extension requires reengineering the $FOLD$ operation to be done on the fly, similar to the process of delayed composition. It would also be interesting to merge this framework with the approach of Wang et al. [14] to fuse DNN and GMM acoustic models.

[1] The reported factor includes computation times for both acoustic model (GMM or DNN) and the actual decode computation.

7. References

[1] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[2] I. Shafran, R. Sproat, M. Yarmohammadi, and B. Roark, "Efficient determinization of tagged word lattices using categorial and lexicographic semirings," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 283–288.

[3] R. van Dalen, A. Ragni, and M. Gales, "Efficient decoding with continuous rational kernels using the expectation semiring," Citeseer, Tech. Rep., 2012.

[4] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[5] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, vol. 27. Baltimore, 2000.

[6] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in hmm/ann multi-stream asr," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–741.

[7] A. Janin, D. P. Ellis, and N. Morgan, "Multi-stream speech recognition: Ready for prime time?" in *Eurospeech 99: 6th European Conference on Speech Communication and Technology: Budapest, Hungary, September 5-9, 1999*. ESCA, 1999.

[8] H. Hermansky, "Multistream recognition of speech: Dealing with unknown unknowns," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1076–1088, 2013.

[9] S. Y. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition." in *INTERSPEECH*. Citeseer, 2008, pp. 898–901.

[10] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[11] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[12] X. Li, R. Singh, and R. M. Stern, "Lattice combination for improved speech recognition," in *Proc. of International Conference on Spoken Language Processing*, 2002.

[13] N. Mesgarani, S. Thomas, and H. Hermansky, "Toward optimizing stream fusion in multistream recognition of speech," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. EL14–EL18, 2011.

[14] H. Wang, A. Ragni, M. J. Gales, K. M. Knill, P. C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Proc. Interspeech*, vol. 15, 2015.

[15] W. Lee, J. Kim, and I. Lane, "Multi-stream combination for lvcsr and keyword search on gpu-accelerated platforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3296–3300.

[16] I. L. Hetherington, H. Shu, and J. R. Glass, "Flexible multi-stream framework for speech recognition using multi-tape finite-state transducers," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[17] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[18] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second chimespeech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 126–130.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.