FLEXIBLE MULTI-STREAM FRAMEWORK FOR SPEECH RECOGNITION USING MULTI-TAPE FINITE-STATE TRANSDUCERS

I. Lee Hetherington, Han Shu, and James R. Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{*ilh, hshu, jrg*}*@csail.mit.edu*

ABSTRACT

We present an approach to general multi-stream recognition utilizing multi-tape finite-state transducers (FSTs). The approach is novel in that each of the multiple "streams" of features can represent either a sequence (e.g., fixed- or variable-rate frames) or a directed acyclic graph (e.g., containing hypothesized phonetic segmentations). Each transition of the multi-tape FST specifies the models to be applied to each stream and the degree of feature stream asynchrony to allow. We show how this framework can easily represent the 2-stream variable-rate landmark and segment modeling utilized by our baseline SUMMIT speech recognizer. We present experiments merging standard hidden Markov models (HMMs) with landmark models on the Wall Street Journal speech recognition task, and find that some degree of asynchrony can be critical when combining different types of models. We also present experiments performing audio-visual speech recognition on the AV-TIMIT task.

1. INTRODUCTION

Most commonly, speech recognition systems utilize a single stream of features, often a fixed-rate sequence of observation vectors (e.g., MFCCs and their derivates) modeled using hidden Markov models (HMMs). Extensions to this traditional HMM approach include segmental (e.g., whole-phone) modeling [1, 2], multi-stream sub-band modeling [3], multi-stream multi-rate modeling [4, 5], articulatory-inspired modeling [6, 7], multi-modal recognition [8], and audio-visual speech recognition [9, 10, 11], among others. Many of these multi-stream approaches vary in how often information from the different streams is integrated (e.g., every state, every phone or syllable boundary, or at the end of the utterance) and whether the initial search utilizes all streams or rather additional streams are integrated in a multi-pass approach.

Our SUMMIT speech recognition system [2] has long integrated two feature streams, *landmarks* and *segments*, and integrated them at phone boundaries in an integrated search. At such phone boundaries, determined automatically by the search, the landmark and segment feature streams are fully synchronized in time. When performing some initial experiments combining a traditional HMM with our landmark and segment models, we found that synchronization with the HMM was an issue. We found that our landmark/segment system preferred different phone boundaries as compared to a context-dependent HMM, and thus desired a framework to explore asynchrony in addition to multiple feature streams. Others have found that context-dependent HMMs prefer phonetic alignments that may not well match transcriptions or other models, including context-independent HMMs [12], and thus allowing some degree of asynchrony between HMMs and other models may be critical to successful integration. In this paper we present our multi-stream framework that utilizes a multi-tape finite-state transducer (FST) to express how multiple feature streams are combined and the allowable asynchrony between them at different parts of the search.

Related work includes multi-stream recognition by HMM recombination by Bourlard, Dupont, et al. [3, 9], in which HMMs representing different streams are allowed to evolve independently until encountering special synchronization states. The multi-rate HMM framework of Çetin and Ostendorf [5] utilizes graphical models and allows different streams to operate at different rates. The multi-modal approach of Johnston and Bangalore et al. [8] jointly recognizes gestures and speech using multi-tape FSTs, with integration of the modalities occurring at the end of the utterance (either two passes or search through recognition lattices computed on each modality).

In Section 2 we start with background on our pre-existing 2-stream system and present our new multi-stream framework. In Section 3 we report on experiments run with the new framework, including integration with traditional HMM models and audio-visual speech recognition.

2. MULTI-STREAM, MULTI-TAPE FST FRAMEWORK

In this section we begin with a description of the 2-stream modeling of landmarks and segments utilized by our baseline speech recognizer and then generalize this to arbitrary feature streams allowing asynchrony using a multi-tape FST representation.

2.1. Landmark & Segment Modeling: 2 Streams

Our baseline speech recognizer [2] has long made use of both landmark and segmental acoustic features. Landmarks are proposed with the goal of having them occur at phone boundaries. Segments are proposed with the goal of having them span whole phones. In practice, both landmarks and segments are over-generated, allowing the recognition search to choose the optimal phonetic segmentation. For landmarks this means that some will be proposed internal to phones. For segments this means that a directed acyclic graph is proposed to cover all hypothesized segmentations of the utterance into phones.

The landmark models and the segment models operate on separate feature streams that are both derived from the same set of fixed-rate (5ms) MFCC features. The landmark feature stream is

Support for this research was provided in part by the National Science Foundation under grant #IIS-0415865.

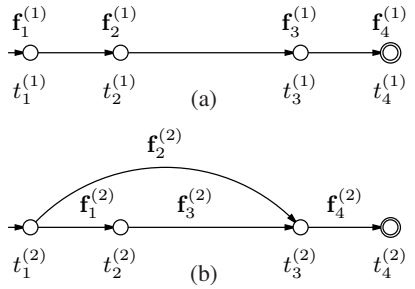

Fig. 1. Sample 2-stream feature space. Stream 1 in (a) consists of variable-rate landmark features, with a feature vector $\mathbf{f}_i^{(1)}$ and time $t_i^{(1)}$ associated with each landmark i. Stream 2 in (b) consists of segments connecting pairs of landmarks, with a feature vector $\mathbf{f}_j^{(2)}$ associated with each segment j and time $t_k^{(2)}$ associated with each segment boundary k.

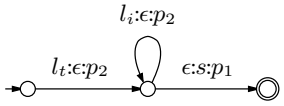

Fig. 2. Topology of 2-stream landmark/segment phonetic model. This combines a 2-state landmark HMM (l_t for transition model and l_i for internal model) operating on stream 1 with a whole-segment model s operating on stream 2.

in general a variable-rate sequence of fixed-length feature vectors, much like those used in a variable-rate HMM [5]. The segment feature "stream" isn't a stream per se but rather a directed acyclic graph, in which each arc represents a time range with its own fixed-length feature vector. Thus, this system operates on two streams, one a variable-rate sequence of landmarks and the other a variable-rate graph of segments. Figure 1 displays an example of two time-aligned feature streams, with the top showing landmarks and the bottom showing segments.

In the landmark feature stream, each phone is modeled with a "transition" model l_t at the phone-initial landmark, and an "internal" model l_i at each phone-internal landmark, if any. In the segment feature stream, each phone is modeled by a single whole-phone model operating on the fixed-length feature vector. We make use of "antiphone" models to account for the overlapping segments [2]. In general, the landmark models (transition and internal) and segment models are all context-dependent. At phone boundaries, the landmark and segment feature streams are fully synchronized in time (i.e., $t^{(1)} = t^{(2)}$).

2.2. New Framework

The modeling outlined in this previous section, particularly the relationship between segments and landmarks, is hardwired in our baseline system. We now present a generalization of this baseline model allowing arbitrary feature streams and asynchrony between them in an integrated search.

In general, we will have F feature streams: each could be a graph containing potential time transitions. At any given point in the search, we will be at some F-tuple $(t^{(1)}, t^{(2)}, \ldots, t^{(F)})$ representing the current time (or state) across all feature streams. We call such a time F-tuple a "hypertime" $\mathbf{t}$.

To do this, we make use of a multi-tape FST representation of our search space. In addition to encoding weights or probabilities for different pronunciations and word sequences, it also encodes allowable transitions through the F feature streams, which models to apply to each, and when the search synchronizes the features streams. We have chosen to encode these multi-stream search characteristics into an FST. For F streams, a transition label will be of the form

$$m^{(1)} : m^{(2)} : \ldots : m^{(F)} : p : o \;/\; w \;.$$

Here, o and w represent word labels and weights (e.g., $-$log probabilities), respectively.[1] Each $m^{(f)}$ represents a model identifier for feature stream f, or ϵ if there is none. p identifies a predicate to be applied to the hypertime $\mathbf{t}$, controlling the degree of asynchrony (in time, as opposed to in states) between the feature streams at any given point in the search, or it too can be ϵ for no predicate.

To make a transition within the FST and hypertime space from $\mathbf{t}_1 \rightarrow \mathbf{t}_2$, the presence of the $m^{(f)}$ model identifiers constrains the possible hypertime transitions. If $m^{(f)} \neq \epsilon$ (i.e., model present for stream f), then feature space f must make a transition: $t_2^{(f)} \neq t_1^{(f)}$. Otherwise, there will be no transition: $t_2^{(f)} = t_1^{(f)}$. In addition, the predicate p on the transition must evaluate to true for the destination hypertime, $p(\mathbf{t}_2) = 1$ if a joint FST and feature space transition is to take place.[2]

When an FST transition is taken, the score is updated by linear combination of the log probabilities provided by the individual feature classifiers. Note that this is a substantial difference from Bourlard et al.'s HMM recombination framework [3] in which the stream scores are integrated only at synchronization states. Earlier integration has the potential advantage that the different streams can contribute to beam pruning earlier in the search, although it does limit the form of feature score combination possible.

We use a dynamic-programming search to find the best path through the FST and through the F feature spaces. We perform a "time"-synchronous beam search starting at the initial hypertime and proceed to the final hypertime, visiting intermediate hypertimes in order.[3] At each hypertime, we perform score-based and count-based beam pruning of the active nodes (FST states). In addition, we can perform beam pruning across all hypertimes that share the same $t^{(1)}$, generally the feature space with the finest time resolution. Ignoring the synchronization predicates, this search could be viewed as computing the best path(s) through a multi-tape join operation [13] between the multi-tape FST above and the multi-tape cross product of feature stream FSTs, where each feature stream FST contains transitions with a model label and model score (i.e., the weight).

In our initial implementation, we have been training the models for individual streams separately and then combining them in this multi-stream framework. Any transition weights used for the individual models are linearly combined in the same way individual model scores are combined. We believe it would be straightforward to train the models jointly as others have done in multi-stream systems.

2.3. Example: Landmarks & Segments

Figure 2 shows the FST representing a phone model implementing the type of modeling used by the baseline recognizer. Tape 1 represents models for the landmark feature stream, tape 2 models for

[1] Figures in this paper have been simplified and do not show the o and w components on FSTs.

[2] An obvious extension would be to use a probabilistic predicate.

[3] The particular ordering used is not important. We used lexicographically sorted order for convenience.

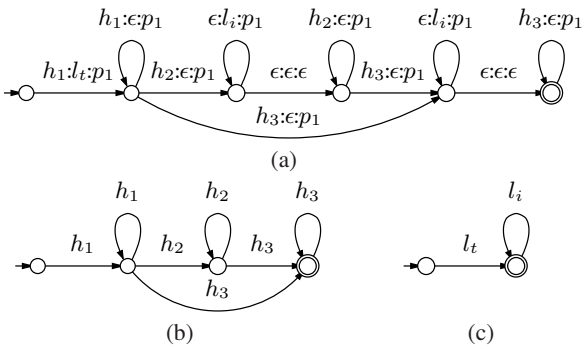

Fig. 3. (a) FST jointly modeling fixed-rate frames on stream 1 and variable-rate landmarks on frame 2. It is a combination of the individual stream FSTs shown in (b) and (c). Other constructions are possible, including a full Cartesian product of the state spaces (b) and (c), but (a) is the topology used in our experiments.

the segment feature stream, and tape 3 references to hypertime predicates.

Predicate p_1 enforces the degree of asynchrony allowed at phone boundaries: $p_1(\mathbf{t}) = |t^{(1)} - t^{(2)}| \leq \tau$. For the baseline system $\tau = 0$, but in this framework p_1 allows us to relax this synchrony constraint. Predicate p_2 is used to keep feature stream 1 from getting too far ahead of stream 2: $p_2(\mathbf{t}) = t^{(1)} \leq \max\text{reachable}(t^{(2)}) + \tau$, where "max reachable" represents the maximum finishing time of a segment starting at $t^{(2)}$, and improves efficiency by eliminating dead ends in the search.

2.4. Example: Frames & Landmarks

Figure 3 shows an FST representing a 2-stream model combining a 3-state fixed-rate frame HMM and a 2-state variable-rate landmark HMM. In this case we have interleaved the two models to produce the 2-stream FST. This FST allows partial-phone frame and landmark scores to be integrated early and compete within the beam search. Other topologies are possible, including the full Cartesian product of the individual stream FSTs. However, unless models are sensitive to the state of other streams (e.g., landmark model used depends on state of frame HMM), such an expansion offers no advantage within our framework. The synchronization predicate is $p_1(\mathbf{t}) = |t^{(1)} - t^{(2)}| \leq \tau$, meaning the frame and landmark streams are allowed to be out of sync by up to τ, both between and within phones. As demonstrated in the next section, we have found $\tau = 95$ms to work well.

3. EXPERIMENTS

We have experimented with the multi-stream framework on two different tasks. One is the Wall Street Journal (WSJ) speech recognition task, and the other is an audio visual speech recognition task on the AV-TIMIT corpus [14].

3.1. The WSJ Task

The WSJ corpus consists of read speech of sentences from the *Wall Street Journal* newspaper. We chose to do the standard H2-C2 task on the Eval'92 test set [15, 16]. For training, we used the WSJ SI84 corpus. The training set contains 14 hours of speech with 7,138 sentences. The language model is a bigram with a decoding vocabulary of 5,000 words. The Eval'92 test set has 330 sentences with 5,353 words with 0.29% OOV rate.

Acoustic Models	Test WER
Landmark Models	10.4%
HMM Models	8.8%
Landmark + HMM Models	8.0%

Table 1. Word error rates (WER) for variable-rate landmark models, fixed-frame-rate HMMs, and their combined models.

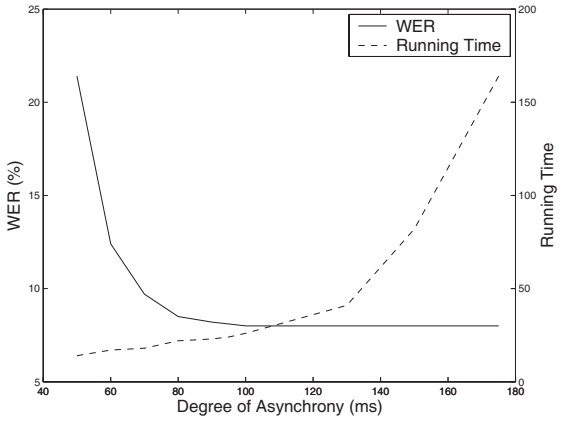

Fig. 4. WERs and decode time vs. degree of asynchrony.

Two baseline systems were used for this WSJ task. The first baseline system is a standard HMM system. The 42-dimensional feature vector consists of 14-dimensional MFCCs, their deltas, and their delta-deltas, all computed with a fixed 10ms frame shift. The 3-state HMM acoustic models have 3,347 clustered triphone models with 26,742 gaussians. The word error rate (WER) for the baseline HMM system is 8.8%. The second baseline system uses a 50-dimensional landmark feature vector computed at hypothesized landmark locations. The landmark features can be thought of as having a variable frame rate and contain various MFCC averages on both sides of the landmark. The average landmark spacing is approximately 30ms. The baseline landmark acoustic models had 993 clustered diphone models with 13,496 gaussians. The WER using the landmark models is 10.4%.

The multi-stream decoder provides a flexible framework to combine these two baseline feature streams. The multi-tape FST representation of the phone model used for these two streams is the same one illustrated in Figure 3(a).

Table 1 summarizes the results in terms of WERs of the baseline landmark models, HMM models, and their combined models. The combined acoustic models achieved a WER error rate of 8.0%, which improves from either baseline models alone. A development set was used to optimize the weighting of the landmark and HMM scores.

The degree of asynchrony allowed in the time predicates has a significant impact on performance in terms of WER and computation time. In general the landmarks and the HMM features are not aligned in time, and a strict requirement of all the phonetic boundaries are synchronized at the same locations will mostly likely not produce any complete hypothesis. Figure 4 shows how the WER and the computation time changes as the degree of asynchrony varies. When the asynchrony between the two streams is at least 95ms ($\tau \geq 95$ms), the WER does not improve and the computation time in-

Acoustic Models	Test WER
Speech Landmark & Segment	2.27%
Visual HMM Models	96.3%
Speech Landmark & Segment + Visual HMM Models	0.91%

Table 2. WERs for speech landmark and segment models, visual HMMs, and their combined models.

creases. The two feature stream need a minimum amount of asynchrony so a compatible hypothesis can be considered. The computation time increases with increasing degree of asynchrony because the size of search space increases due to an increasing number of hypertimes visited.

3.2. The AV-TIMIT Task

The multi-stream framework can also be applied to other recognition tasks. Here we show its use for audio-visual speech recognition on the AV-TIMIT corpus. The AV-TIMIT corpus is a collection of audio-visual speech data of many speakers reading phonetically rich TIMIT sentences. Along with the speech waveform, the facial movement of the speakers were also captured in video. The training set consists of 3,608 utterances from 185 speakers, and the test set contains 285 utterances from other 19 speakers [14].

Two baseline systems were used for this task. The first baseline system was a segment-based system using both landmark and segment features from audio data only. The second baseline system was a frame-based 3-state HMM system modeling only the visual features. The multi-stream system modeled these three feature streams together.

Table 2 summarizes the results in terms of WERs of the baseline speech landmark and segment models, visual HMM models, and their combined models. The WER is the same as reported in [17], where a custom designed decoder was used for combining the same 3 feature streams. The decoding speeds of the two decoders were approximately the same.

4. DISCUSSION & FUTURE WORK

The work reported in this paper summarizes our effort to construct a new flexible framework for multi-stream speech recognition. The framework takes advantage of the multi-tape FST representation to provide flexibility in representation. The individual feature streams in this framework can include, for example, fixed-rate or variable-rate sequences of frames or more generally directed acyclic graphs such as phonetic segments. When operating with a single stream, the framework can represent a traditional frame-based HMM or a segment-based system. When operating with multiple streams, it enables the integration of diverse feature streams, potentially improving overall system accuracy. In this paper, we demonstrated two multi-stream systems: one a combination of a variable-rate landmark model with a standard HMM for the WSJ task and the other a combination of a landmark model, a segment model, and a visual HMM for the AV-TIMIT task.

With the specification of the time predicates on the multi-tape FST, the framework also provides the means to accommodate the possible asynchrony among the various feature streams. Because the degree of asynchrony allowed, as expressed by time predicates, we have found that good design of the predicates themselves can be important in minimizing decoding time.

In this paper, we experimented with combining two or three feature streams. In future, we plan to combine more feature streams, such as sub-band features or articulatory features, but, in general the search space is exponential in the number of feature streams. If decoding becomes too computationally demanding, a multi-pass approach might be desirable. Here we might perform a first pass integrating a subset of the streams, and then use the output of this pass (e.g., word or phone lattice) to constrain a second pass integrating the reamainder of the streams. To date, we have focused primarily on synchronization at the phonetic level, but we wish to extend the approach to the syllable level in the future.

5. ACKNOWLEDGMENTS

We would like to thank T. J. Hazen for his help with AV-TIMIT and his 3-stream (landmark, segment, and visual) recognition system.

6. REFERENCES

[1] M. Ostendorf, V. Digilakis, and O. Kimball, "From HMMs to segment models: a unified view of stochastic modelling for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[2] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, no. 2, pp. 137–152, 2003.

[3] H. Bourlard, S. Dupont, and C. Ris, "Multi-stream speech recognition," Tech. Rep. IDIAP-RR 96-07, IDIAP, 1996.

[4] S. Dupont and H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system," in *Proc. Eurospeech*, Rhodes, Sept. 1997, pp. 3–6.

[5] Ö. Çetin, *Multi-Rate Modeling, Model Inference, and Estimation for Statistical Clusters*, Ph.D. thesis, University of Washington, Seattle, 2004.

[6] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2000.

[7] K. Livescu and J. R. Glass, "Feature-based pronunciation modeling with trainable synchrony probabilities," in *Proc. ICSLP*, Jeju, South Korea, Oct. 2004.

[8] M. Johnston and S. Bangalore, "Finite-state multimodal parsing and understanding," in *Proc. 18th Conf. on Computational Linguistics*, Saarbrücken, 2000, pp. 369–375.

[9] S. Dupont and J. Luettin, "Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database," in *Proc. ICSLP*, Sydney, Nov. 1998, pp. 1283–1286.

[10] C. Chibelushi, F. Deravi, and J. Mason, "A review of speech-based bimodal recognition," vol. 4, no. 1, pp. 23–37, Mar. 2002.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," vol. 91, no. 9, pp. 1306–1326, Sept. 2003.

[12] D. T. Toledano, L. A. Hernández Gómez, and L. V. Grande, "Automatic phonetic segmentation," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, Nov. 2003.

[13] A. Kempe, J.-M. Champarnaud, J. Eisner, F. Guingne, and F. Nicart, "A class of rational n-WFSM auto-intersections," in *Proc. Conf. Impl. and Appl. of Automata*, Sophia Antipolis, June 2005, pp. 189–200.

[14] T. J. Hazen, K. Saenko, C. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development and initial experiments," in *Proc. ICMI*, State College, Pennsylvania, Oct. 2004.

[15] F. Kubala et al., "The hub and spoke paradigm for CSR evaluation," in *Proc. ARPA Human Language Technology Workshop*, Princeton, Mar. 1994, pp. 37–42.

[16] J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, Cambridge, UK, 1995.

[17] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. Acoustics and Audio Processing*, May 2006, to be published.