Evaluation of a WFST-based ASR System for Train Timetable Information

Josef Robert Novak*, Edward W. D. Whittaker† and Sadaoki Furui*
* Department of Computer Science, Tokyo Institute of Technology 152-8552 Tokyo Japan
E-mail: {novakj,furui}@furui.cs.titech.ac.jp
† Inferret Japan 101-0047 Tokyo Japan
E-mail: ed@inferret.jp

Abstract---**This paper focuses on the description and evaluation of our new prototype Weighted Finite State Transducer (WFST) based ASR system for accessing Japanese train timetable information from a mobile device. The system employs a constrained, WFST grammar which permits users to make natural language queries related to the train timetables domain, including queries related to supplementary concepts such as ``last train", ``arrival time", etc. We report preliminary results for four related WFST networks, which include weighted and un-weighted variants of an Zenkoku network which covers 9199 stations across Japan, and weighted and un-weighted variants of a Tokyo network which contains 1098 stations from the greater Tokyo area. For weights we employ unigram station probabilities which were estimated from search engine log data. We compare the performance of the un-weighted networks to their weighted counterparts, using a test set which includes 474 utterances produced by 6 different speakers, and which was recorded in a variety of different acoustic environments. In particular, we focus on task completion rates, including accuracy on at-least-one-station, both-stations, and both-stations-plus-time. In the best case, for the 10-best, weighted Zenkoku network we report a both-stations accuracy of 92.2%.**

I. Introduction

Automated speech driven services, for mobile and otherwise, have been a popular area of research for some time. Two examples of such services include the JUPITER system [1], introduced by researchers at MIT in 1997, and the How-May-I-Help-You system [2], which was deployed by AT&T in 2001. The JUPITER system allows users to access worldwide weather information from a telephone using spoken dialogue, while How-May-I-Help-You employs speech recognition and spoken language understanding technology to route calls based on un-restricted user input.

The system which we are currently developing focuses on information related to train timetables, specifically with regard to the Japanese rail system. The train timetables domain has also received some research interest in the past, specifically in [3] which describes a telephony-based dialogue system for fielding queries about German intercity train timetables, and [4] which describes a speech recognition enabled multimodal system for querying Dutch train timetables. We have also developed a similar application in the past which is described in [5].

In contrast to [3] and [4], our current system employs a one-shot, question-answering (QA) style approach which encourages users to query the system using natural language questions rather than using keywords and which involves only a single step, as opposed to a multi-step interactive dialogue. In contrast to our own previous work in [5] which focused on ASR over the telephony network, the current system employs an internet-based client-server architecture, where audio is recorded locally on the mobile device via an embedded application and subsequently sent to the server where speech recognition is actually performed. The results are then returned to the user.

The primary focus of the experiments we report is task completion with respect to station name recognition accuracy in spoken queries, independent of supplemental time information such as ``last train", ``arrival time", etc. In particular we look at the effectiveness of applying a unigram model based on station frequency data, and compare this to a baseline system which does not employ any form of language model or weighting scheme.

Evaluation results are based on 898 spoken utterances, half of which were used to tune the recognition system configuration, and half of which were used for testing. The utterances were recorded by a total of 6 different speakers in a variety of different acoustic environments ranging from a quiet room to a station platform. Users were given minimal directions on how to use the system and encouraged to experiment with different speaking strategies.

The remainder of the paper is structured as follows, Section II. provides a brief outline of the WFST speech recognition framework, Section III. provides a description of the experimental setup, Section IV. describes the experimental results. Section V. provides further discussion of the experimental results, including some discussion of observed ASR errors. Section VI. concludes the paper and discusses future work.

II. WFST-based Speech Recognition

Throughout this paper we employ the Weighted Finite State Transducer (WFST) approach to speech recognition. The WFST approach has gained considerable popularity in recent years as it provides an elegant, unified framework for handling a wide variety of speech recognition problems. In particular, this framework facilitates optimization and pre-compilation of the recognition network which can often lead to reduced processing time. The WFST approach also facilitates easy integration and modification of component knowledge sources [6].

The speech recognition setup we employ for these experiments focuses on a transducer cascade which maps from context-dependent triphones to word sequences, that is $C \circ L \circ G$. Here C represents the Context-dependency transducer, L represents the Lexicon transducer, and G represents the Grammar or language model transducer.

For further technical details and mathematical motivations behind the WFST approach please see [7].

III. Experimental Setup

In this section we describe some salient aspects of our experimental setup, including acoustic and language model characteristics, and tools used to construct and optimize the WFST recognition network.

The acoustic models used in these experiments were trained with SphinxTrain using the Corpus of Spontaneous Japanese (CSJ) [8] and all audio data was resampled to 8khz prior to training. The acoustic models employed a three state left-to-right topology, and the state output densities were 16 component Gausian mixture models.

The recognition grammar used for the experiments consisted of a hand-written expert grammar which was based on observations on a set of approximately 100 spoken train-timetables queries, which were not used in the experiments. A simplified regular-expression based syntax was employed to specify the grammar, and a custom tool was written to convert the resulting expressions to equivalent NFAs. The most popular or most requested verbal strategies for asking questions about supplementary time-related concepts such as ``departure time", ``arrival time", ``last train", ``first train" were incorporated into the network along with the global requirement that the ``to station", and ``from station" be supplied prior to any supplementary concepts. Thus an example of a valid spoken query might be,

「六本木から上野駅まで、07:30出発」
``I want to go from Roppongi to Ueno station, leaving at 07:30."

While an example of an unsupported query might be,

「10:30までに着きたい、広尾から恵比寿まで」
``I want to arrive at 10:30, I'm going from Hiro to Ebisu."

Purely keyword-based queries were also allowed e.g.,

「自由ヶ丘 津田沼 10:30 到着」
``Jiyuugaoka Tsudanuma 10:30 arrival"

was also considered a valid query and permitted by the grammar.

In order to evaluate the effect of station vocabulary size on recognition quality, two separate station vocabularies were prepared. The larger vocabulary, **Zenkoku**, contained 9199 stations from the Japan national rail system, while the smaller vocabulary, **Tokyo**, contained 1098 stations, all from the greater Tokyo area.

Apart from the size of the station vocabulary the **Zenkoku** network and the **Tokyo** network were identical. In addition to the un-weighted **Zenkoku** and **Tokyo** networks, weighted variants were also constructed.

Unigram weights for the weighted version of the **Zenkoku** network, **Zenkoku-uni**, and the weighted version of the **Tokyo** network, **Tokyo-uni**, were estimated from text-based mobile search logs which were kindly provided by an industry company, and incorporated directly into the WFST grammar. These unigram weights were calculated using add-one smoothing,

$$P_{+1}(e) = \log \frac{c(e)+1}{N+V}$$

where $c(e)$ refers to the raw number of times station e occurred, as either an origin or a destination, $N = 1029270$ refers to the total number of unigram counts for all stations e in the vocabulary, and V refers to the vocabulary size. No weights were estimated or explicitly applied to any other parts of the grammar. Thus in total, four different networks were constructed which differed based on station vocabulary size and weighted versus un-weighted. The characteristics of the WFST networks used in the experiments are summarized in Table I.

TABLE I
Network Characteristics.

Network	Vocab size	Weights
Tokyo	1098	None
Tokyo-uni	1098	Unigram
Zenkoku	9199	None
Zenkoku-uni	9199	Unigram

The open source OpenFST [9] library and associated command line tools were used to construct, compose and optimize the networks used in these experiments.

IV. Test Data

The audio data used for the experiments consisted of 948 spoken utterances, which were split into a development set, and an evaluation set. Utterances were recorded using a web-based client server system where the client took the form of an iPhone application. Audio data was recorded locally on the client in .wav format, using an 8khz sampling rate, then transferred to the server where speech recognition was actually performed. The iPhone application was installed on 4 iPhones, and users were allowed to query the system freely, at any time, following very limited formal instruction, aside from a cursory explanation of the grammar limitations. The audio data included utterances from a total of 6 different speakers, including 2 females and 4 males.

The development set, which consisted of 474 of the recorded utterances, was used to tune the decoder configuration, the parameters of which included an insertion penalty, a language model weight, and two parameters which are used to specify hypothesis pruning thresholds.

The evaluation set, which consisted of the remaining 474 utterances, was used for testing, and results reported in the following sections are based on this data set.

V. Experimental Results

In this section we report results from our experiments. We focus on task completion rates in the form of recognition accuracy for either-station (**either**), both-stations (**both**), and both-stations-plus-time (**all**).

Text-based mobile as well as PC oriented train timetables search engines typically provide a backoff strategy in the form of drop-down lists in order to better handle ambiguous user input. The natural extension to this in the speech recognition case is an N-best list or confusion networks.

In our tests we also extracted the 10-best station candidates for the ``from station'' and the ``to station'' from the recognition lattices generated by the decoder.

These results are illustrated in Table II.

TABLE II
Task completion rates.

Network	either (10-best)	both (10-best)	all (10-best)
Tokyo	96.6% (98.9%)	71.9% (87.5%)	63.7% (77.4%)
Zenkoku	86.9% (97.1%)	48.1% (70.7%)	42.6% (62.7%)
Tokyo-uni	99.2% (99.8%)	78.7% (92.2%)	69.8% (81.6%)
Zenkoku-uni	97.9% (99.7%)	68.4% (90.5%)	61.2% (80.4%)

In particular Table II. shows the task completion rates for the four networks in the **both** case, when only the 1-best result is considered, and when the 10-best results are considered. The relative improvement, **rel**, is also displayed. Figure 1. shows the task completion rates for the **both** case for the various networks as a histogram.

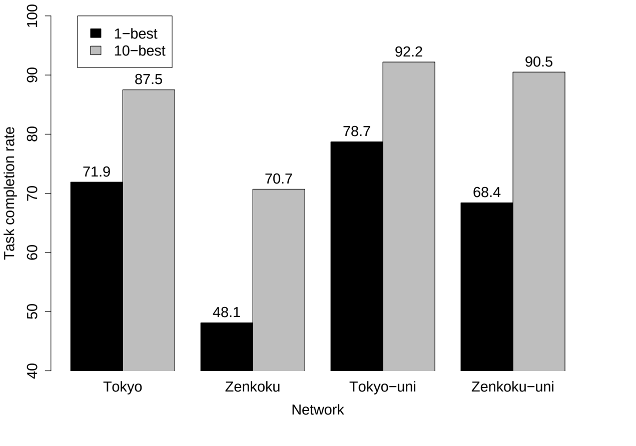

Fig. 1. Task completion rates.

Finally, in order to evaluate the significance of our results, we performed a McNemar test comparing the 10-best results of the **Tokyo** and **Tokyo-uni**, and **Zenkoku** and **Zenkoku-uni** networks, respectively. The McNemar test and its motivations are described in detail in [10].

TABLE III
Distribution of Errors for **Tokyo**, and **Tokyo-uni**.

		Tokyo-uni	
		Correct	Incorrect
Tokyo	Correct	883	1
	Incorrect	27	37

Table IV. illustrates the distribution of errors with regard to the **Tokyo** and **Tokyo-uni** networks.

We calculate the chi-squared value for Table IV. using the standard measure,

$$X^2 = \frac{(|U-W|-1)^2}{U+W} \ ,$$

where U refers to the errors made only by the un-weighted network, and W refers to errors made only by the weighted network. This yields a chi-squared value of $X^2 = 22.3$ and a corresponding P-value $P < 0.0001$.

Table V. shows the corresponding distribution of errors for the **Zenkoku** and **Zenkoku-uni** networks,

TABLE IV
Distribution of Errors for **Zenkoku**, and **Zenkoku-uni**.

		Zenkoku-uni	
		Correct	Incorrect
Zenkoku	Correct	793	2
	Incorrect	109	44

which yield a corresponding chi-squared value of $X^2 = 101.2$. Here again the resulting P-value $P < 0.0001$. In both cases the P-value is considerably less than 0.0001, thus the results of the McNemar tests indicate that in both cases, the improvements in task-completion rates gained from the application of the unigram model can be considered statistically significant with high confidence.

VI. Discussion

In the previous section we provided several task completion results based on the accuracy of our train timetables ASR system. In this section we provide further discussion of these results.

First, from Table II. it can be clearly seen that the **Zenkoku** network performs the worst out of the group which is hardly surprising as the **Zenkoku** network contains the largest number of stations, and no weights. It can also be seen from Table II. that task completion rates for the **all** case are consistently lower than the task completion rates for the **either** and the **both** (stations only) cases.

Figure 1. clearly shows, as expected, that employing 10-best results for the stations provides a large boost to performance. In particular, in the case of the **Zenkoku** network it can be seen that simply providing the 10-best candidates for each station yields a relative improvement in the **both** case, of 43.5%. Similarly, in the case of the **Zenkoku-uni** network, we see a very large relative improvement of 69.9%.

Furthermore the results of the McNemar tests indicate that the improvements we achieve are statistically significant with high confidence. It is worth noting however, that the relatively small size of the test set and its focus on the Tokyo area stations, may also have introduced some positive bias to the improvements in the **Zenkoku** case.

A. Discussion of Errors

It is also worth looking at the kinds of errors which the system made. Many of the station name recognition errors could be attributed to users employing speaking strategies which were not supported by the present grammar. Another frequent and obvious source of errors was noise. Recordings which were made on station platforms, or recorded on the street generally yielded a greater number of errors. Another common source of errors originated from utterances where the speaker mis-pronounced or repeated himself, or stuttered. Finally, the user interface for the embedded iPhone application seemed to present a challenge to some users, who consistently started speaking before the recording started, or presumably continued to speak, even after they had released the record button.

VII. Conclusions

In this paper we described preliminary results for a new WFST-based, mobile-oriented ASR system for querying Japanese train timetables. The system employs a domain-restricted, ``one shot'' WFST grammar which permits users to make natural-langauge queries related to the train timetables domain, including queries related to concepts such as ``last train'', ``first train'', or to optionally specify additional information such as ``arrival time'', ``departure time'', etc.

In particular we reported results for four different recognition networks, which differed in terms of station vocabulary size and un-weighted versus weighted, where the weighted case employed unigram weights estimated from search engine logs obtained from a Japanese mobile train timetables search engine.

We showed that, as expected, including unigram weights improves task completion rates in this application, and that the best overall task completion rate could be achieved by further using the N-best results. In the best case for the **Zenkoku-uni** network, we showed an 10-best task completion rate for the **both** case of 90.5%, while in the best case for the **Tokyo-uni** network, we showed a task completion rate of 92.2%. Although in live, telephony-based speech recognition applications using N-best results is often impractical due to the time required to playback or synthesize the individual recognition candidates, in a multi-modal environment such as that supported by the iPhone and similar devices it is a simple matter to leverage these results through use of a touch screen. In future we plan to evaluate the effect of N-best lists in this application on overall user satisfaction. Finally, we employed a standard McNemar significance test to our results, in order to evaluate the significance of the improvement gained through the application of the unigram station weights. The results of this test were encouraging and indicate, with $P < 0.0001$, that the observed improvement is highly unlikely to represent a chance occurrence.

In these experiments we employed an elementary unigram language model, however employing more sophisticated techniques will almost certainly further improve the results. In particular we plan to investigate the use of bigram station priors, and more sophisticated smoothing methods.

In the train timetables domain it is conceivable that station priors change based on time of day, therefore it may also prove interesting to evaluate the effect of using priors based on different periods of the day. In recent years GPS functionality has become commonplace in many Japanese cellular phones and mobile devices, thus it may also be interesting to investigate the effect of conditioning the ASR results on the user's current location.

Finally, in these experiments we employed a fairly restrictive grammar, however in future we plan to also evaluate performance employing less restricted models such as a large vocabulary trigram model.

Acknowledgment

We would like to thank Inferret Japan for providing access to their WFST decoder and data collection framework, and Gourmet Navigator Inc. for providing us with the search engine log data used to compute the station priors.

References

[1] V. Zue, S. Sene, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, ``Telephone-based Conversational Interface for Weather Information,'' *IEEE Trans. on Speech and Audio Processing*, vol. 8, 2000.

[2] A.L. Gorin, B.A. Parker, R.M. Sachs, and J.G. Wilpon ``How May I Help You?,'' *IVTTA* 1996.

[3] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, E.G. Schukat-talamazzini, ``A Spoken Dialogue System for German Intercity Train Timetable Inquiries,'' *In Proc. European Conf. on Speech Technology*, pp. 1871-1874, 1993.

[4] J. Sturm, I. Bakx, B. Cranen, J. Terken, F. Wang, ``Usability Evaluation of a Dutch Multimodal System for Train Timetable Information,'' *Proc. LREC2002, Gran Canaria de Las*, 2002.

[5] E.W.D. Whittaker, P. Dixon, J. Novak, S. Furui, ``A Prototype Spoken Natural Language Interface for Information Access on Mobile Phones,'' *In Proc. Acoustic Society of Japan*, 2007.

[6] P.R. Dixon, D.A. Caseiro, T. Oonishi, and S. Furui, ``The TITech large vocabulary WFST speech recognition system,'' *Proc. Automatic Speech Recognition and Understanding*, Kyoto, Japan, pp.443-448 (2007-12).

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

[8] K. Maekawa, ``Corpus of Spontaneous Japanese Its Design and Evaluation,'' *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7-12, 2003.

[9] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut and M. Mohri, ``OpenFst: A General and Efficient Weighted Finite-State Transducer Library,'' *Proc. of the Ninth International Conference on Implementation and Application of Automata*, (CIAA 2007), volume 4783 of Lecture Notes in Computer Science, pages 11-23. Springer, 2007.

[10] L. Gillick, S. Cox, ``Some Statistical Issues in the Comparison of Speech Recognition Algorithms,'' *In Proc. ICASSP*, 1989.