FAST WORD-GRAPH GENERATION FOR SPONTANEOUS CONVERSATIONAL SPEECH TRANSLATION

Tohru Shimizu, Harald Singer and Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Labs.
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
{shimizu,singer,sagisaka}@itl.atr.co.jp

ABSTRACT

This paper introduces the latest advances in research at ATR on speech translation for spontaneous conversations, especially focusing on speech recognition efforts. For recognition, we employ a word search technique that generates moderate sized word graphs in real-time. To cope with a variety in length of utterances, e.g., word, phrase, sentence fragment, sentence, and concatenated sentences in spontaneous speech, we have adopted a two pass search strategy that uses variable-order word n-gram statistics in the first stage and task dependent language constraints in the second stage. This strategy is evaluated using the "ATR Travel Arrangement" corpus.

1. INTRODUCTION

Currently, a next-generation speech translation system that aims for natural trans-language communication is under development at ATR. This system recognizes Japanese conversational speech, translates it into English, German and Korean, and outputs synthesized speech for each language. To cope with spontaneous speech phenomena such as filled pauses, hesitations and corrections, a large number of word (sentence) hypotheses have to be considered during the speech recognition process. Since the conventional CFG based one-pass search strategy (HMM-LR) adopted in our previous speech translation system (ASURA) [1], required much computation to manage both the acoustic state and syntactic state, especially for long sentences or ill-formed sentences, a recognition scheme using word graphs has been applied in our new speech recognition system.

In Section 2, the target domain of our speech translation research and characteristics of the data collected are described. In Section 3, the configuration of the speech-translation system, which adopts a word graph as an interface between speech processing and linguistic processing, is described. Next, in Sections 4 and 5, a recognizer overview and the method for fast word graph generation are given. Finally, we present current experimental results obtained using the "ATR Travel Arrangement" corpus.

2. TRAVEL ARRANGEMENT TASK

At ATR, a speech-translation system is under development using the "ATR Travel Arrangement" corpus [2, 3]. This task provides an image of a useful application for people who have felt a real need for automatic speech translation on some occasions, such as when conversing with a travel agent or hotel staff in a foreign country.

The database used for speech recognition research is composed of two parts: an integrated speech and language database (bilingual) and a speech database (monolingual). The integrated speech and language database includes conversations that take place between native Japanese and English speakers through human interpreters. This database has been designed for a moderate degree of spontaneity and a fairly large-sized (-10^4) vocabulary. The speech database, on the other hand, is designed to cover speaker variations and high spontaneity.

Table 1 and Figure 1 show the database size used for developing the speech recognizer (English data is shown for reference) and vocabulary growth of the bilingual database, respectively. The vocabulary size for 150,000 spoken words is 5,405 (English) and 4,526 (Japanese), compared to 2,000 (English) for the "Scheduling Task", which has been collected at CMU for spontaneous speech translation research [4]. It is interesting that the "Hotel Reservation Task", which is a subset of the "Travel Arrangement Task" and accounts for 19% of the conversations, requires a larger vocabulary than the "Scheduling Task" of CMU.

The characteristics for spontaneous Japanese speech are listed in Table 2. The spontaneity of the bilingual speech data has been affected by constraints associated with speaking through interpreters.

Table 1. Size of database used for system development

	dialogues	words	speakers
Japanese			
(bilingual)	618	240,522	52
(monolingual)	228	74,533	143
English			
(bilingual)	618	202,746	-

Table 2. Number of filled pauses, corrections and false starts per utterance (Japanese customer utterance only)

	filled pauses	corrections, false starts
bilingual	0.29	0.006
monolingual	0.46	0.039

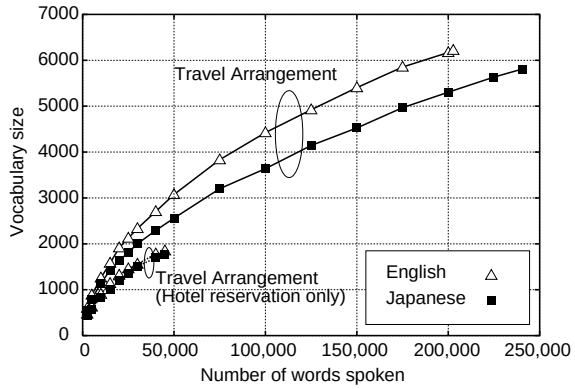

Figure 1. Comparison of vocabulary growth between English and Japanese (bilingual database)

3. SYSTEM OVERVIEW

The current speech translation system consists of 1) a spontaneous continuous speech recognizer that generates a compact word graph (ATRSPREC) [5], 2) a transfer-driven machine translator that uses high-speed translation-example retrieval and incremental parsing (TDMT) [6], 3) a multilingual speech re-sequencing synthesis system (CHATR) [7], and 4) a lattice rescoring and sub-lattice generation procedure, which supplies a word graph with reasonable size to be processed in the translation stage. In all of these four components, statistical methods are employed to cope with the huge number of word hypotheses resulting from highly coarticulated spontaneous speech and differences in speaking conditions between training and testing. A word graph data structure has been adopted as an interface between speech processing and linguistic processing to transfer multiple sentence hypotheses in a compact form (Figure 2).

The word graph format is based on the Standard Lattice Format in HTK version 2.0 (SLF), enhanced in a number of ways to deal with online speech translation. For example, allowing multiple graphs in a stream, outputting absolute time (wall clock time) at the start of the utterance to implement a multi-modal speech interface, emitting a finite state automaton (FSA) node label for the simultaneous use of FSA and n-gram, discriminating between phonetically identical words (homonyms), and outputting the acoustic model name (when multiple competing acoustic models are used) to synthesize translated speech using speaker characteristics as close as possible to the input speaker.

4. RECOGNIZER OVERVIEW

In continuous speech recognition, especially during a fast search for real-time implementation, reducing the number of word (sentence) hypotheses is a crucial issue. To reduce both the processing time and the time delay between the end of an utterance and the recognition result output, most of the computation in our multi-pass approach is concentrated in the first pass. Precise acoustic models and language models are applied in the first pass, and a language score is re-evaluated to reduce the overall graph size in the second pass.

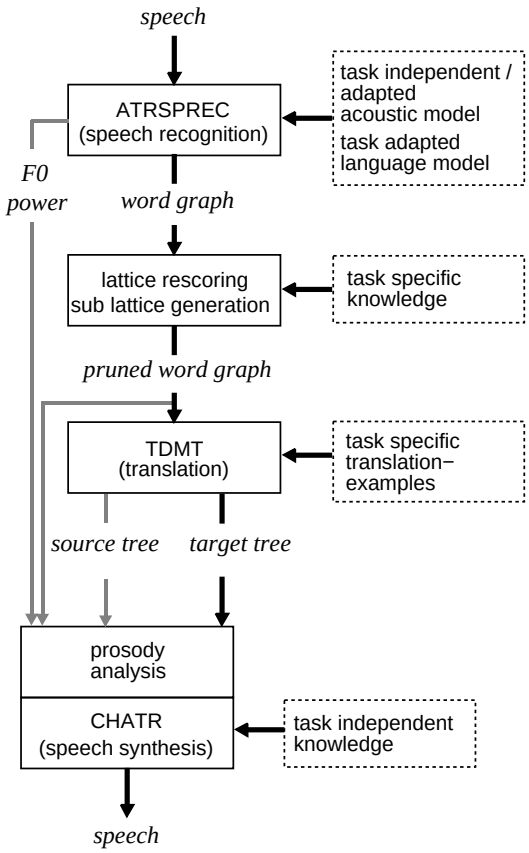

Figure 2. Current configuration of the ATR speech translation system

Acoustic model: The acoustic model is a state shared phoneme HMM (HMnet), which is a contextual dependent model with a small number of parameters generated by using the maximum likelihood-based successive state splitting (ML-SSS) algorithm [8]. To quickly adapt parameters using a small amount of input speech (by balancing the reliability of original parameters in existing models and newly estimated parameters), MAP-VFS speaker adaptation [9] is applied.

Language model: The language model is a variable-order n-gram that provides reliable statistical constraints from a given language corpus with fewer parameters than conventional n-grams. A variable-order n-gram language model is generated using a word-class splitting and consecutive word grouping algorithm [10].

Lexicon: The lexicon is represented phonemically using a set of 26 phonemes and allows multiple pronunciations. The lexicon includes both predefined words and consecutive word sequences, which are generated during the process of variable order n-gram model generation.

Decoding: The recognizer uses a time-synchronous beam search based fast word graph generation algorithm that includes both intra-word and inter-word context dependent phone models, and time-asynchronous graph search for rescoring the graph [5].

5. FAST WORD GRAPH GENERATION

We define a word hypothesis in the word graph to be (t, t_b, s, h, c, V, $Q_A(t, t_b, s, h, c)$, $Q_L(s, v)$, $Q_P(t_b, v)$),
where

t:	time
t_b:	start time (word)
s:	lexicon node
h:	HMM state at lexicon node s
c:	allophone of word head
V:	set of preceding words
v:	preceding word !Jv ∈ V !K
$Q_A(t, t_b, s, h, c)$:	acoustic likelihood at t, t_b!$ s!$ h!$ c
$Q_L(s, v)$:	language likelihood at s, v
$Q_P(t_b, v)$:	accumulated likelihood from start of utterance at t_b!$ v.

Word hypothesis merging and pruning have to be considered to reduce the computational cost of the acoustic likelihood ($Q_A(t, t_b, s, h, c)$) and language likelihood ($Q_L(s, v)$) calculations.

5.1. Cross-word context approximation

Word hypothesis pruning, which determines t_b from several possibilities, is well known as a "word-pair approximation" [11]. This pruning has the effect of reducing the acoustic likelihood ($Q_A(t, t_b, s, h, c)$) calculation. However, as the "word-pair approximation" uses the preceding word when determining the word boundary t_b, the number of preceding words for each word hypothesis is restricted to one (size V is equal to one), even though many preceding words have the same word ending portion (t_b might be the same for words sharing the same word end pronunciation, e.g., in English hotel and tell).

To share word hypotheses as much as possible (size V could be larger than one), we use "cross-word context" as the preceding word information ("cross-word context approximation") [5].

5.2. Language score look-ahead and tying

Language score look-ahead techniques have been proposed to apply language likelihoods as early as possible in a tree lexicon [12, 13]. The estimated language likelihood of the word hypotheses sharing the same lexicon node s is derived considering the probabilities of all possible word continuations E_s. This technique has an effect on word hypothesis merging, since word hypotheses that have the same initial phone sequence share the same language likelihood ($Q_L(s, v)$). Another advantage of this method is that the correct language likelihood of word w is applied at the word end lexicon node $S(w)$.

$$Q_L(s, v) = \max_u P(u|v) \quad (1)$$

where $u \in E_s$.

$$Q_L(S(w), v) = P(w|v) \quad (2)$$

However, even when language score look-ahead is applied, the computational cost is still huge, as short length homonyms, which match local characteristics of the speech, frequently appear and have different language likelihoods. In our implementation, homonyms are treated as one word in the first pass, and they are separated into individual words and language likelihood is re-evaluated in the second pass. The language likelihood applied in the first pass for homonym w is the maximum likelihood of all members of the homonym class $C(w)$.

$$Q_L(S(w), v) = \max_{k,l} P(w_k|w_l) \quad (3)$$

where $w_k, w \in C(w), w_l, v \in C(v)$.

5.3. Language score interpolation between lexicon nodes

Undesirable pruning in the beam search may occur when language likelihood is close to the pruning threshold. To reduce these pruning errors, changes of the log likelihood between succeeding lexicon nodes are linearly interpolated [5].

6. SPEECH RECOGNITION RESULTS

As the word graph output from the speech recognizer is re-evaluated in a lattice rescoring process (see Figure 2), it is important that the correct words are included in the graph, but not necessarily on the highest scoring path.

In our evaluation, word graph density and two recognition rates "rank1" (word recognition rate of the highest likelihood path), "max" (word recognition rate of the highest recognition rate path, i.e., upper bound in the word graph) are used. Word graph density that indicates the size of a word graph is defined as follows.

$$word\ graph\ density = \frac{number\ of\ word\ hypotheses}{number\ of\ spoken\ words} \quad (4)$$

As a test set, 7 dialogues (3 male, 4 female) were selected from the integrated speech and language database. A 6.6K lexicon, which includes the vocabulary of the whole task (utterances of customer, clerk and interpreter), and a 1.3K lexicon, which includes vocabulary from the customer's utterances in the "Hotel Reservation Task", were used for evaluation. Other experimental conditions are summarized in Table 3.

Figure 3 gives the results for the test set. The "max" word recognition rate for the 1.3K lexicon and 6.6K lexicon were 81.6% and 75.1%, respectively. Since the difference between "max" and "rank1" was about 15% for a reasonable sized word graph, the word graph data structure seems to provide an efficient interface. This is especially true when model parameters are not reliable, e.g., when speech is highly coarticulated or when speaking conditions between training and testing are not matched.

Table 4 shows the effect of the approximation methods described in Section 5. The approximation methods significantly reduce cpu-time. All tests were performed on a HP 9000/735 workstation (135Specint92).

7. SUMMARY

This paper has focused on speech recognition algorithms within the framework of ATR's multi-lingual speech-to-speech translation system, which is currently under development.

Table 3. Experimental conditions

Analysis conditions	
Sampling rate	12 kHz
Window	Hamming window (20 ms)
Frame period	10 ms
Analysis	log power + 16-order LPC-Cep + Δlog power + 16-order ΔLPC-Cep
Acoustic model (HMnet)	
Topology	401 states, 5 mixtures
Training	2,620 words
Retraining	150 sentences (read speech)
Adaptation (speaking-style) (speaker)	 128 utterances (non-read speech) 1 dialogue (non-read speech)
Language model (variable-order n-gram)	
Training	308,518 words (828 dialogues)
Number of classes	713
Word perplexity	49.6

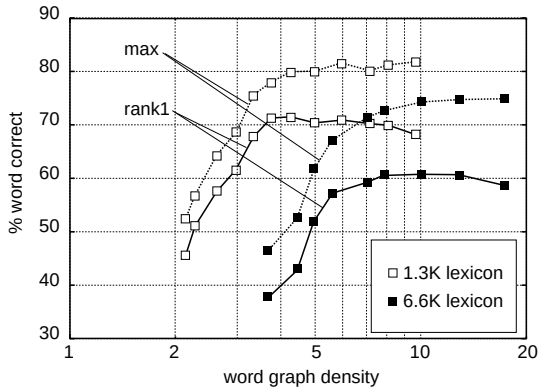

Figure 3. Word recognition rate for various word graph sizes

This system uses context-dependent state sharing HMM's and variable-order n-gram statistics trained by using a travel arrangement dialogue corpus. A fast word graph generation method that allows almost real-time recognition on a 135Specint92 workstation has been developed. Recognition performance has been evaluated for test sets with vocabularies of up to 6,600 words.

Currently, a number of techniques, such as discriminative metric design for feature extraction [14], detailed acoustic modeling using segment models and stochastic pronunciation networks [15], recurrent neural networks [16], and incremental adaptation [17], are being studied and we plan to incorporate them into our future system.

ACKNOWLEDGMENTS

The authors would like to thank Hirokazu Masataki for supplying a variable-order n-gram language model. We would also like to thank Hirofumi Yamamoto for supporting experiments.

Table 4. Effect of cpu-time reduction by applying approximations (6.6K lexicon, cpu-time to achieve 72% word recognition rate (max))

approximation method	cpu-time (x real-time)
look-ahead	117.8
look-ahead, tying	2.8
look-ahead, tying, cross-word context approximation	1.6
look-ahead, tying, cross-word context approximation, language score interpolation	1.2

REFERENCES

[1] T. Morimoto et al.: "ATR's Speech Translation System: ASURA," Proc. of Eurospeech'93, pp. 1291-1294, (1993).

[2] T. Morimoto et al.: "A Speech and Language Database for Speech Translation Research," Proc. of ICSLP'94, pp. 1791-1794, (1994).

[3] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka: "Japanese Speech Databases for Robust Speech Recognition," Proc. of ICSLP'96, pp. 2199-2202, (1996).

[4] A. Waibel et al.: "JANUS-II - Translation of Spontaneous Conversational Speech," Proc. of ICASSP'96, pp. 409-412, (1996).

[5] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka: "Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graphs," Proc. of ICASSP'96, pp. 145-149, (1996).

[6] O. Furuse, J. Kawai, H. Iida, S. Akamine and D. Kim: "Multi-Lingual Spoken-Language Translation Utilizing Translation Examples," Proc. of NLPRS'95, pp. 544-549, (1995).

[7] N. Campbell: "CHATR: A High-Definition Speech Re-Sequencing System," Proc. of ASA and ASJ Third Joint Meeting, pp. 1223-1228, (1996).

[8] H. Singer and M. Ostendorf: "Maximum Likelihood Successive State Splitting," Proc. of ICASSP'96, pp. 601-604, (1996).

[9] M. Tonomura, T. Kosaka and S. Matsunaga: "Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Probability Estimation," Journal of Computer Speech and Language, vol. 10, pp. 117-132, (1996).

[10] H. Masataki and Y. Sagisaka: "Variable-Order N-gram Generation by Word-Class Splitting and Consecutive Word Grouping," Proc. of ICASSP'96, pp. 188-191, (1996).

[11] H. Ney and X. Aubert: "A Word Graph Algorithm for Large Vocabulary, Continuous Speech Recognition," Proc. of ICSLP'94, pp. 1355-1358, (1994).

[12] V. Steinbiss, B.H. Tran and H. Ney: "Improvements in Beam Search," Proc. of ICSLP'94, pp. 2143-2146, (1994).

[13] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young: "The 1994 HTK Large Vocabulary Speech Recognition System," Proc. of ICASSP'95, pp. 73-76, (1995).

[14] H. Watanabe, T. Yamaguchi and S. Katagiri: "Discriminative Metric Design for Pattern Recognition," Proc. of ICASSP'95, pp. 3439-3442, (1995).

[15] T. Fukada, M. Bacchiani, K. Paliwal and Y. Sagisaka: "Speech Recognition Based on Acoustically Derived Segment Units," Proc. of ICSLP'96, pp. 1077-1080, (1996).

[16] M. Schuster: "Learning Out of Time Series with an Extended Recurrent Neural Network," Proc. of IEEE Neural Networks for Signal Processing'96, pp. 170-179, (1996).

[17] Q. Huo and C.H. Lee: "A Study of On-Line Quasi-Bayes Adaptation of CDHMM-Based Speech Recognition," Proc. of ICASSP'96, pp. 705-708, (1996).

FAST WORD-GRAPH GENERATION FOR SPONTANEOUS CONVERSATIONAL SPEECH TRANSLATION

Tohru Shimizu, Harald Singer and Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Labs.
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
{shimizu,singer,sagisaka}@itl.atr.co.jp

This paper introduces the latest advances in research at ATR on speech translation for spontaneous conversations, especially focusing on speech recognition efforts. For recognition, we employ a word search technique that generates moderate sized word graphs in real-time. To cope with a variety in length of utterances, e.g., word, phrase, sentence fragment, sentence, and concatenated sentences in spontaneous speech, we have adopted a two pass search strategy that uses variable-order word n-gram statistics in the first stage and task dependent language constraints in the second stage. This strategy is evaluated using the "ATR Travel Arrangement" corpus.