SPEECH-TO-SPEECH TRANSLATION BASED ON FINITE-STATE TRANSDUCERS *

F. Casacuberta[1] *D. Llorens*[2], *C. Martínez*[1], *S. Molau*[3], *F. Nevado*[1], *H. Ney*[3],
M. Pastor[1], *D. Picó*[1], *A. Sanchis*[1], *E. Vidal*[1], *J. M. Vilar*[2]

[1] Dpt. de Sistemes Informàtics i Computació, Institut Tecnològic d'Informàtica
Universitat Politècnica de València, 46071 Valencia, SPAIN.
[2] Dpt. d'Informàtica, Universitat Jaume I, 12071 Castelló de la Plana, SPAIN
[3] Lehrstuhl für Informatik VI, RWTH Aachen
University of Technology, D-52056 Aachen, GERMANY

ABSTRACT

Nowadays, the most successful speech recognition systems are based on stochastic finite-state networks (hidden Markov models and n-grams). Speech translation can be accomplished in a similar way as speech recognition. Stochastic finite-state transducers, which are specific stochastic finite-state networks, have proved very adequate for translation modeling. In this work a speech-to-speech translation system, the EUTRANS system, is presented. The acoustic, language and translation models are finite-state networks that are automatically learnt from training samples. This system was assessed in a series of translation experiments from Spanish to English and from Italian to English in an application involving the interaction (by telephone) of a customer with a receptionist at the front-desk of a hotel.

1. INTRODUCTION

The present and most successful speech recognition systems are based on the *integrated* application of an acoustic and a language model. This integration can be carried out due to the use of finite-state networks as models, particularly Hidden Markov Models (HMM) for acoustic modeling and n-grams or stochastic finite-state grammars for language modeling. However, the present and most common speech-to-speech translation systems are based on a *serial* architecture composed by a speech recognition system followed by a linguistic (or more recently, statistical [1]) text-to-text translation system.

The possibility of using stochastic finite-state networks for limited-domain translation has been discussed in previous works [2, 3]. These models obviously support the above mentioned conventional *serial* architecture. More interestingly, these models are also quite adequate for a fully *embedded* architecture, where the acoustic models are *integrated* into the translation model in a similar way as for speech recognition. In any case, due to the finite-state nature of all the involved models, the procedure for translation in the serial architecture, or for integrated recognition-and-translation in the embedded architecture, are based on the very same *Viterbi* search engine.

One of the main objectives of the EUTRANS project was the development of machine translation systems for limited-domain tasks with speech input [4]. Of particular interest were finite-state transducers, which can be built automatically from examples. EUTRANS was a five-year joint effort of four European institutions (ITI in València/Spain, RWTH in Aachen/Germany, FUB in Rome/Italy and ZERES GmbH in Bochum/Germany), partially funded by the *Open Domain* of the *Long-Term Research (LTR)* ESPRIT program of the European Union.

In this paper, we present the speech translation prototypes that have been built using the methodologies developed and the data collected in the EUTRANS project.

2. FINITE-STATE TRANSDUCERS AND SPEECH TRANSLATION

Let x be the acoustic representation of an input sentence. The translation of x into another language can be formulated as the problem of searching for a sequence of words $\hat{s}$ in the target language that maximize

$$\hat{s} = \operatorname*{argmax}_{s} \Pr(s|x). \qquad (1)$$

But translation can be also seen as a two steps process

$$x \rightarrow e \rightarrow s,$$

where e is a possible decoding of x in the source language that can be translated into a sequence of words, s, in the

*THIS WORK HAS BEEN PARTIALLY SUPPORTED BY THE EUROPEAN UNION UNDER GRANT IT-LTR-OS-30268.

target language. With the assumption that $\Pr(x|e,s)$ does not depend on the output s, and using the max-operator as an approximation to the sum,

$$\hat{s} \approx \operatorname*{argmax}_{s} \max_{e} \Pr(e,s) \cdot \Pr(x|e). \qquad (2)$$

In practice, $\Pr(x|e)$ is modeled by acoustic models (HMMs) and $\Pr(e,s)$ by a translation model which, in our case, is a stochastic finite-state transducer.

A *stochastic finite-state transducer* is a finite-state network whose transitions are labeled by three items: (i) an input symbol (a word from the input vocabulary), (ii) an output string (a sequence of words from the output vocabulary) and (iii) a transition probability. Fig. 1 shows a small fragment of a stochastic finite-state transducer for Italian to English translation.

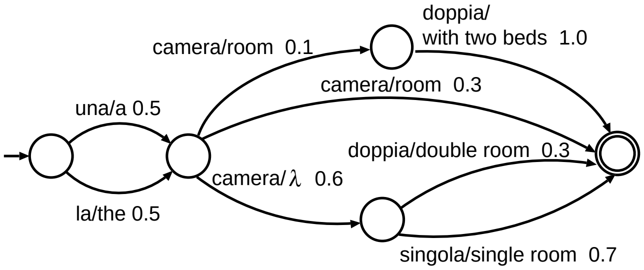

Fig. 1. Example of a stochastic finite-state transducer. "λ" denotes the empty string. The input sentence *"una camera doppia"* can be translated to either *"a double room"* or *"a room with two beds"*. The most probable translation is the first one with a probability of 0.09.

With such a translation model, the probability $\Pr(e,s)$ can be easily computed by summing up the probabilities of all paths that account for the translation pair (e,s). The probability of each path is the product of the probabilities of the transitions involved in the path.

The maximisation of Eq. 2 can be performed by a search process in an *integrated network*, similar to the ones used for speech recognition. Each transition of the stochastic finite-state transducer is expanded into the concatenation of HMMs of the phone units that define the input word of the transition.

As in the case of standard speech recognition, in general, this search process is a difficult computational problem [5]. Nevertheless, quite adequate approximations can be obtained by using the Viterbi algorithm on a *trellis* associated to the input acoustic sequence and the integrated network.

Obviously, for translating a given acoustic sequence x, a *serial architecture* can also be straightforward implemented using stochastic finite-state transducers. Eq. 2 becomes

$$\hat{s} \approx \operatorname*{argmax}_{s} \max_{e} \Pr(s|e) \cdot \Pr(e) \cdot \Pr(x|e).$$

The maximisation can be approximated as follows:

1. *Word decoding* of x by searching for a sequence of words $\hat{e}$ such that

   $$\hat{e} \approx \operatorname*{argmax}_{e} \Pr(e) \cdot \Pr(x|e),$$

   where $\Pr(x|e)$ is modeled by an acoustic and $\Pr(e)$ a language model.

2. Given $\hat{e}$, the *translation* of $\hat{e}$ by searching for a sequence of words $\hat{s}$ such that

   $$\hat{s} \approx \operatorname*{argmax}_{s} \Pr(s|\hat{e}) = \operatorname*{argmax}_{s} \Pr(\hat{e},s),$$

   where $\Pr(\hat{e})$ is assumed to be independent of s, and $\Pr(\hat{e},s)$ is modeled by a stochastic finite-state transducer.

3. THE EUTRANS SYSTEM

One of the most important aims of the EUTRANS project was to develop machine translation systems to assist human to human (speech) communications [4]. More specifically, a telephone speech input translation system, capable to translate telephone calls from one language into another has been developed.

This system is based on the ATROS (Automatically Trainable Recognizer Of Speech) engine. ATROS is a continuous-speech recognition/translation system which uses stochastic finite-state models at all its levels: acoustic-phonetic, lexical and syntactic/translation. All these models can be obtained in an automatic way. This makes the system easily adaptable to different recognition/translation tasks. A first version of ATROS for Spanish continuous speech recognition was presented in [6].

The ATROS system is completely coded in the C programming language. It only needs a general-purpose CPU without the help of any digital signal processor. This allows a great portability and hardware independence. The system currently runs on three different Unix platforms.

3.1. Acoustic, lexical and translation models

During signal analysis, short-term spectral analysis is performed on short overlapping signal segments (*frames*). The resulting power spectrum is warped according to the Mel-scale, a filterbank is applied, and cepstral coefficients are derived from the log filterbank outputs.

The different knowledge sources used in ATROS are:

- **Acoustic-Phonetic models**: Each (context-dependent) phoneme (including silence) is described by a continuous Gaussian mixture density HMM [7].

- **Lexical models**: Each word is represented by a stochastic finite state automaton, automatically generated from the allophonic description of the word.

- **Translation models**: Stochastic finite state transducers are built automatically from a training corpus of paired sentences [8, 9].

3.2. Linguistic/translation decoding

The translation procedure of the ATROS system is based on a Viterbi beam-search for the optimal path in a finite-state network which integrates all the above mentioned models.

The translation of an input sentence is built by concatenating the output strings of the successive transitions that compose the optimal path.

3.3. Speech-Input translation prototypes

Two speech-to-speech translation prototypes have been implemented, one for Spanish to English and the other for Italian to English. In both cases, the general application was the translation of queries, requests and complains made by telephone to the front desk of a hotel. However, the Italian-English task was significantly more complex and closer to a real situation than the Spanish-English one.

The output English speech is obtained by using a free software Text-To-Speech synthesizer which offers understandable speech at reasonably good quality.

EUTRANS-I: speech-input Spanish-English translation

This system is fully operational for both telephone and microphone input.

The acoustic models of phone units were left-to-right continuous-density HMMs. They were trained with the HTK Toolkit [10]. 26 Spanish monophones were used. The translation model was trained with the OMEGA transducer inference algorithm [8].

The text corpus was generated in a semi-automatic way using travel booklets as a seed corpus. From a selected subset of these text data, a multi-speaker Spanish speech corpus was produced. The utterances were acquired using both a microphone and a telephone.

The text training corpus was composed of 10,000 pairs of sentences (132,198/ 134,922 running Spanish/ English words). The size of the Spanish/English vocabularies were 686/513 and the corresponding bigram test set perplexities were 8.6/ 5.2, respectively. The speech corpus for training (Spanish) phone HMMs was composed of 11,000 running words. The speech test set consisted of 336 Spanish sentences (3,000 running words). It should be noted that this corpus is significantly smaller than the overall corpus produced in the project and used in [2, 3]. Recently, the mentioned subset of 10,000 training pairs was established as a more realistic training corpus for the kind of application considered.

EUTRANS: speech-input Italian-English translation

This second system is also fully operational through standard telephone lines for remote (or local) operation.

The acoustic models of phone units were also left-to-right continuous density HMMs. These models were trained using a Viterbi approximation [7]. Best performance was obtained with decision-tree clustered generalized triphones (CART with 1,500 tied states plus silence). A linear discriminant analysis (LDA) further improved the recognition accuracy.

The translation model was trained with the "Morphic Generator Translation Inference" technique introduced in [9].

The speech corpus consisted of acquisitions of real phone calls to the front desk of a hotel, simulated using *Wizard of Oz* techniques [11]. This corpus is highly spontaneous and contains many non-speech artifacts. The text corpus was obtained by manually transcribing the acquired Italian utterances and translating them into corresponding English sentences.

From this text corpus, 3,038 pairs of sentences (61,423/ 72,689 running Italian/English words) were used for training the translation model. The Italian/English vocabularies had 2,459/1,701 words and the corresponding bigram test set perplexities were 31/25, respectively. The speech training corpus used to train the (Italian) phone models was composed of 52,511 running words. The speech test set consisted of 278 Italian sentences (5,381 running words).

3.4. Prototype assessment

To assess the performance of the systems, two error criteria were used. On the one hand, the *(Recognition) Word Error Rate (WER)* for the decoding of the speech input. On the other hand, the *Translation Word Error Rate (TWER)*, i.e. the WER obtained by comparing each automatically translated sentence with a *single reference target sentence*. Because most source-language sentences allow for many correct target translations, TWER should be considered a pessimistic error estimation. This is particularly true in the case of the Italian-English application due to the free-form human-produced test set reference translations.

The Italian-English EUTRANS prototype achieves quite acceptable response time (about three times real time or less), while the Spanish-English EUTRANS -I prototype often runs in less than real time, even on low-cost Pentium machines.

Assessment results of the EUTRANS -I prototype are presented in Table 1.

Table 1. Assessment results of the EuTrans -I prototype (Tel = telephone, Mic = microphone)

Models and conditions			WER(%)	TWER(%)
OMEGA	Tel	integrated	12.8	15.4
OMEGA	Tel	serial 3-Gr	11.1	14.1
OMEGA	Mic	integrated	5.1	6.8
OMEGA	Mic	serial 3-Gr	4.7	6.8

In the original experiments with the EuTrans -I prototype [3], the results with microphone input in the integrated architecture were better than those reported here. The main reason was that in those experiments a huge amount of training data was used to train the finite-state transducers. In the present experiment, in order to approach more realistic conditions, the amount of training data was dramatically reduced.

Assessment results of the EuTrans prototype are presented in Table 2. Using transducers trained with OMEGA the results achieved were clearly worse than the ones achieved by MGTI and are not reported here for the sake of brevity.

Table 2. Assessment results of the EuTrans prototype.

Models		WER(%)	TWER(%)
MGTI	serial 3-Gr	22.1	37.9
MGTI	integrated	32.0	44.8

Both for EuTrans -I and for EuTrans , the results presented with a serial architecture were achieved by using a trigram language model for the input speech decoding.

4. DISCUSSION AND CONCLUSIONS

Two prototypes have been implemented for speech-to-speech translation. One for translation from Italian to English and another for translation from Spanish to English. Both support all kinds of finite-state translation models. They run on low-cost hardware and are fully accessible through standard telephone lines. Response times are close to or better than real time.

From our present results, it appears that the integrated recognition/translation architecture performs similar or worse than the serial coupling of a speech recognizer and a translation module. We believe, however, that this is caused by insufficient training data for transducer learning, as suggested by our previous results with simpler tasks and/or larger training sets [2, 3].

5. REFERENCES

[1] H. Ney; S. Nießen; F. Och; H. Sawaf; C. Tillmann; S. Vogel, “Algorithms for statistical translation of spoken language,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 24–36, 2000.

[2] E.Vidal, “Finite-state speech-to-speech translation,” in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 1997, pp. 111–114.

[3] J.C. Amengual; J.M. Benedí; F. Casacuberta; A. Casta no; A. Castellanos; V.M. Jiménez; D. Llorens; A. Marzal; M. Pastor; F. Prat; E. Vidal; J.M. Vilar, “The EUTRANS-I speech translation system,” *Machine Translation Journal*, to appear 2001.

[4] Instituto Tecnológico de Informática; Fondazione Ugo Bordoni; Rheinisch Westfälische Technische Hochschule Aachen; Lehrstuhl für Informatik VI; Zeres GmbH Bochum, “Example-based language translation systems. Final report,” Tech. Rep., Information Technology. Long Term Research Domain. Open scheme. Project Number 32026, 2000.

[5] F.Casacuberta; C. de la Higuera, “Computational complexity of problems on probabilistic grammars and transducers,” in *Grammatical Inference: Algorithms and Applications*, vol. 1891 of *Lecture Notes in Artificial Intelligence*, pp. 15–24. Springer-Verlag, 2000.

[6] D. Llorens; F. Casacuberta; E. Segarra; J.A. Sánchez; P. Aibar, “Acustical and syntactical modeling in ATROS system,” in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 1999, pp. 641–644.

[7] H. Ney; L. Welling; S. Ortmanns; K. Beulen; F. Wessel, “The RWTH large vocabulary continuous speech recognition system,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 853–856.

[8] J.M.Vilar, “Improve the learning of subsequential transducers by using alignments and dictionaries,” in *Grammatical Inference: Algorithms and Applications*, vol. 1891 of *Lenture Notes in Artificial Intelligence*, pp. 298–312. Springer-Verlag, 2000.

[9] F. Casacuberta, “Inference of finite-state transducers by using regular grammars and morphisms,” in *Grammatical Inference: Algorithms and Applications*, vol. 1891 of *Lecture Notes in Artificial Intelligence*, pp. 1–14. Springer-Verlag, 2000.

[10] S. Young; J. Odell; D. Ollason; V. Valtchev; P. Woodland, *The HTK Book (Version 2.1)*, Cambridge University Department and Entropic Research Laboratories Inc., 1997.

[11] D. Aiello; L. Cerrato; C. Delogu; A. Di Carlo, “The acquisition of a speech corpus for limited domain translation,” in *Proceeding of the EUROSPEECH99*, Budapest, 1999.