Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 39-44.
Association for Computational Linguistics.

Architectures for speech-to-speech translation using finite-state models

Francisco Casacuberta **Enrique Vidal**
Dpt. de Sistemes Informàtics i Computació &
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
46071 València, SPAIN.
fcn@iti.upv.es, evidal@iti.upv.es

Juan Miguel Vilar
Dpt. de Llenguatges i Sistemes Informàtics
Universitat Jaume I
Castelló, SPAIN.
jvilar@lsi.uji.es

Abstract

Speech-to-speech translation can be approached using finite state models and several ideas borrowed from automatic speech recognition. The models can be Hidden Markov Models for the accoustic part, language models for the source language and finite state transducers for the transfer between the source and target language. A "serial architecture" would use the Hidden Markov and the language models for recognizing input utterance and the transducer for finding the translation. An "integrated architecture", on the other hand, would integrate all the models in a single network where the search process takes place. The output of this search process is the target word sequence associated to the optimal path. In both architectures, HMMs can be trained from a source-language speech corpus, and the translation model can be learned automatically from a parallel text training corpus. The experiments presented here correspond to speech-input translations from Spanish to English and from Italian to English, in applications involving the interaction (by telephone) of a customer with the front-desk of a hotel.

1 Introduction

Present finite-state technology allows us to build speech-to-speech translation (ST) systems using ideas very similar to those of automatic speech recognition (ASR). In ASR the acoustic hidden Markov models (HMMs) can be integrated into the language model, which is typically a finite-state grammar (e.g. a N-gram). In ST the same HMMs can be integrated in a translation model which consists in a stochastic finite-state transducer (SFST). Thanks to this integration, the translation process can be efficiently performed by searching for an optimal path of states through the integrated network by using well-known optimization procedures such as (beam-search accelerated) Viterbi search. This "integrated architecture" can be compared with the more conventional "serial architecture", where the HMMs, along with a suitable source language model, are used as a front-end to recognize a sequence of source-language words which is then processed by the translation model. A related approach has been proposed in (Bangalore and Ricardi, 2000; Bangalore and Ricardi, 2001).

In any case, a pure pattern-recognition approach can be followed to build the required systems. Acoustic models can be trained from a sufficiently large source-language speech training set, in the very same way as in speech recognition. On the other hand, using adequate learning algorithms (Casacuberta, 2000; Vilar, 2000), the translation model can also be learned from a sufficiently large training set consisting of source-target parallel text.

In this paper, we comment the results obtained using this approach in EUTRANS, a five-year joint effort of four European institutions, partially funded by the European Union.

2 Finite-state transducers and speech translation

The statistical framework allow us to formulate the speech translation problem as follows: Let $\mathbf{x}$ be an acoustic representation of a given utterance; typically a sequence of acoustic vectors or *"frames"*. The translation of $\mathbf{x}$ into a target-language sentence can be formulated as the search for a word sequence, $\hat{\mathbf{t}}$, from the target language such that:

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}|\mathbf{x}). \tag{1}$$

Conceptually, the translation can be viewed as a two-step process (Ney, 1999; Ney et al., 2000):

$$\mathbf{x} \rightarrow \mathbf{s} \rightarrow \mathbf{t},$$

where $\mathbf{s}$ is a sequence of source-language words which would match the observed acoustic sequence $\mathbf{x}$ and $\mathbf{t}$ is a target-language word sequence associated with $\mathbf{s}$. Consequently,

$$\Pr(\mathbf{t}|\mathbf{x}) = \sum_{\mathbf{s}} \Pr(\mathbf{t}, \mathbf{s}|\mathbf{x}), \tag{2}$$

and, with the natural assumption that $\Pr(\mathbf{x}|\mathbf{s}, \mathbf{t})$ does not depend on the target sentence $\mathbf{t}$,

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} \left(\sum_{\mathbf{s}} \Pr(\mathbf{s}, \mathbf{t}) \cdot \Pr(\mathbf{x}|\mathbf{s}) \right). \tag{3}$$

Using a SFST as a model for $\Pr(\mathbf{s}, \mathbf{t})$ and HMMs to model $\Pr(\mathbf{x}|\mathbf{s})$, Eq. 3 is transformed in the optimization problem:

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} \left(\sum_{\mathbf{s}} \Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) \cdot \Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s}) \right), \tag{4}$$

where $\Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t})$ is the probability supplied by the SFST and $\Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s})$ is the density value supplied by the corresponding HMMs associated to $\mathbf{s}$ for the acoustic sequence $\mathbf{x}$.

2.1 Finite-state transducers

A SFST, $\mathcal{T}$, is a tuple $\langle Q, \Sigma, \Delta, R, q_0, F, P\rangle$, where Q is a finite set of *states*; q_0 is the *initial state*; Σ and Δ are finite sets of *input symbols* (source words) and *output symbols* (target words), respectively ($\Sigma \cap \Delta = \emptyset$); R is a *set of transitions* of the form (q, a, ω, q') for $q, q' \in Q$, $a \in \Sigma$, $\omega \in \Delta^\star$ and[1] $P : R \rightarrow \mathbb{R}^+$ (*transition probabilities*) and $F : Q \rightarrow \mathbb{R}^+$ (*final-state probabilities*) are functions such that $\forall q \in Q$:

$$F(q) \quad + \sum_{\substack{\forall (a, \omega, q') \,\in\, \Sigma \times \Delta^\star \times Q \,: \\ (q, a, \omega, q') \,\in\, R}} P(q, a, \omega, q') \quad = \quad 1.$$

Fig. 1 shows a small fragment of a SFST for Spanish to English translation.

A particular case of finite-state transducers are known as subsequential transducers (SSTs). These are finite-state transducers with the restriction of being deterministic (if (q, a, ω, q), $(q, a, \omega', q') \in R$, then $\omega = \omega'$ and $q = q'$). SSTs also have output strings associated to the (final) states. This can fit well under the above formulation by simply adding an end-off-sentence marker to each input sentence.

For a pair $(\mathbf{s}, \mathbf{t}) \in \Sigma^\star \times \Delta^\star$, a *translation form*, ϕ, is a sequence of transitions in a SFST $\mathcal{T}$:

$$\begin{aligned} \phi: \quad & (q_0, \mathrm{s}_1, \tilde{\mathrm{t}}_1, q_1), (q_1, \mathrm{s}_2, \tilde{\mathrm{t}}_2, q_2), \\ & \ldots, (q_{I-1}, \mathrm{s}_I, \tilde{\mathrm{t}}_I \,, q_I), \end{aligned}$$

where $\tilde{\mathrm{t}}_j$ denotes a substring of target words (the empty string for $\tilde{\mathrm{t}}_j$ is also possible), such that $\tilde{\mathrm{t}}_1 \; \tilde{\mathrm{t}}_2 \ldots \tilde{\mathrm{t}}_I = \mathbf{t}$ and I is the length of the source sentence $\mathbf{s}$. The probability of ϕ is

$$\Pr_{\mathcal{T}}(\phi) = F(q_I) \cdot \prod_{i=0}^{I} P(q_{i-1}, \mathrm{s}_i, \tilde{\mathrm{t}}_i, q_i). \tag{5}$$

Finally, the probability of the pair $(\mathbf{s}, \mathbf{t})$ is

$$\begin{aligned} \Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) \quad &= \sum_{\phi \in d(\mathbf{s}, \mathbf{t})} \Pr_{\mathcal{T}}(\phi) && (6) \\ &\approx \max_{\phi \in d(\mathbf{s}, \mathbf{t})} \Pr_{\mathcal{T}}(\phi), && (7) \end{aligned}$$

where $d(\mathbf{s}, \mathbf{t})$ is the set of all translation forms for the pair $(\mathbf{s}, \mathbf{t})$.

These models have implicit source and target language models embedded in their definitions, which are simply the marginal distributions of $\Pr_{\mathcal{T}}$. In practice, the source (target) language model can be obtained by removing the target (source) words from each transition of the model.

[1] By $\Delta^\star$ and $\Sigma^\star$ we denote the sets of finite-length strings on Δ and Σ, respectively

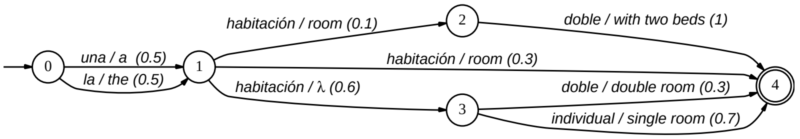


Figure 1: Example of SFST. λ denotes the empty string. The source sentence *"una habitación doble"* can be translated to either *"a double room"* or *"a room with two beds"*. The most probable translation is the first one with probability of 0.09.

The structural (states and transitions) and the probabilistic components of a SFST can be learned automatically from training pairs in a single process using the *MGTI* technique (Casacuberta, 2000). Alternatively, the structural component can be learned using the *OMEGA* technique (Vilar, 2000), while the probabilistic component is estimated in a second step using *maximum likelihood* or other possible criteria (Picó and Casacuberta, 2001). One of the main problems that appear during the learning process is the modelling of events that have not been seen in the training set. This problem can be confronted, in a similar way as in language modelling, by using smoothing techniques in the estimation process of the probabilistic components of the SFST (Llorens, 2000). Alternatively, smoothing can be applied in the process of learning both components (Casacuberta, 2000).

2.2 Architectures for speech translation

Using Eq. 7 as a model for $\Pr(\mathbf{s},\mathbf{t})$ in Eq. 4,

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} \left(\sum_{\mathbf{s}} \max_{\phi \in d(\mathbf{s},\mathbf{t})} \Pr_{\mathcal{T}}(\phi) \cdot \Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s}) \right), \tag{8}$$

For the computation of $\Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s})$ in Eq. 8, let $\mathbf{b}$ be an arbitrary *segmentation* of $\mathbf{x}$ into I acoustic subsequences, each of which associated with a source word (therefore, I is the number of words in $\mathbf{s}$). Then:

$$\Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s}) = \sum_{\mathbf{b}} \prod_{i=1}^{I} \Pr_{\mathcal{M}}(\bar{\mathbf{x}}_i|\mathbf{s}_i), \tag{9}$$

where $\bar{\mathbf{x}}_i$ is the i-th. acoustic segment of $\mathbf{b}$, and each source word $\mathbf{s}_i$ has an associated HMM that supplies the density value $\Pr_{\mathcal{M}}(\bar{\mathbf{x}}_i|\mathbf{s}_i)$.

Finally, by substituting Eq. 5 and Eq. 9 into Eq. 8 and approximating *sums* by *maximisations*:

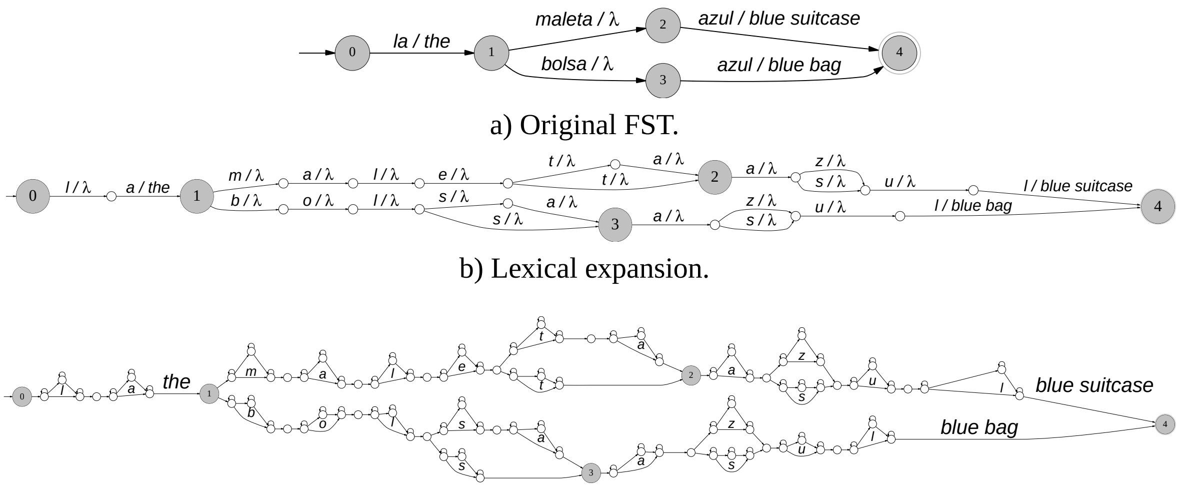
$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\phi \in d(\mathbf{s},\mathbf{t}),\mathbf{b}} \prod_{i=1}^{I} P(q_{i-1}, \mathbf{s}_i, \bar{\mathbf{t}}_i, q_i) \cdot \Pr_{\mathcal{M}}(\bar{\mathbf{x}}_i|\mathbf{s}_i). \tag{10}$$

Solving this maximisation yields (an approximation to) the most likely *target-language sentence* $\hat{\mathbf{t}}$ for the observed *source-language acoustic sequence* $\mathbf{x}$.

This computation can be accomplished using the well known *Viterbi algorithm*. It searches for an optimal sequence of states in an integrated network (*integrated architecture*) which is built by substituting each edge of the SFST by the corresponding HMM of the source word associated to the edge.

This integration process is illustrated in Fig. 2. A small SFST is presented in the first panel (a) of this figure. In panel (b), the source words in each edge are substituted by the corresponding phonetic transcription. In panel (c) each phoneme is substituted by the corresponding HMM of the phone. Clearly, this direct integration approach often results in huge finite-state networks. Correspondingly, a straightforward (dynamic-programming) search for an optimal target sentence may require a prohibitively high computational effort. Fortunately, this computational cost can be dramatically reduced by means of standard heuristic acceleration techniques such as *beam search*.

An alternative, which sacrifices optimality more drastically, is to break the search down into two steps, leading to a so-called *"serial architecture"*. In the first step a conventional source-language speech decoding system (using just a source-language language model) is used to obtain a single (may be multiple) hypothesis for the sequence of uttered words. In the second step, this text sequence is translated into a target-language sentence.



a) Original FST.

b) Lexical expansion.

c) Phonetic expansion.

Figure 2: Example of the integration process of the lexical knowledge (figure b) and the phonetic knowledge (figure c) in a FST (figure a). λ denotes the empty string in panels a and b. In panel c, source symbols are typeset in small fonts, target strings are typeset in large fonts and edges with no symbols denote empty transitions.

Using $\Pr(\mathbf{s},\mathbf{t}) = \Pr(\mathbf{t} \mid \mathbf{s}) \cdot \Pr(\mathbf{s})$ in Eq. 3 and approximating the sum by the maximum, the optimization problem can be presented as

$$(\hat{\mathbf{t}},\hat{\mathbf{s}}) = \operatorname*{argmax}_{\mathbf{t},\mathbf{s}} \left(\Pr(\mathbf{t}|\mathbf{s}) \cdot \Pr(\mathbf{s}) \cdot \Pr(\mathbf{x}|\mathbf{s})\right), \tag{11}$$

and the two-step approximation reduces to

$$\begin{aligned} \hat{\mathbf{s}} &\approx \operatorname*{argmax}_{\mathbf{s}} \{\Pr(\mathbf{s}) \cdot \Pr(\mathbf{x}|\mathbf{s})\}, & (12)\\ \hat{\mathbf{t}} &\approx \operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}|\hat{\mathbf{s}}) & (13)\\ &= \operatorname*{argmax}_{\mathbf{t}} \Pr(\hat{\mathbf{s}},\mathbf{t}). & (14) \end{aligned}$$

In other words, the search for an optimal target-language sentence is now approximated as follows:

1. *Word decoding of* $\mathbf{x}$. A source-language sentence $\hat{\mathbf{s}}$ is searched for using a source language model, $\Pr_{\mathcal{N}}(\mathbf{s})$, for $\Pr(\mathbf{s})$ and the corresponding HMMs, $\Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s})$, to model $\Pr(\mathbf{x}|\mathbf{s})$:

   $$\hat{\mathbf{s}} \approx \operatorname*{argmax}_{\mathbf{s}} \left(\Pr_{\mathcal{N}}(\mathbf{s}) \cdot \Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s})\right).$$

2. *Translation of* $\hat{\mathbf{s}}$. A target-language sentence $\hat{\mathbf{t}}$ is searched for using a SFST, $\Pr_{\mathcal{T}}(\hat{\mathbf{s}},\mathbf{t})$, as a model of $\Pr(\hat{\mathbf{s}},\mathbf{t})$

   $$\hat{\mathbf{t}} \approx \operatorname*{argmax}_{\mathbf{t}} \Pr_{\mathcal{T}}(\hat{\mathbf{s}},\mathbf{t}).$$

A better alternative for this crude "two-step" approach is to use $\Pr(\mathbf{s},\mathbf{t}) = \Pr(\mathbf{s} \mid \mathbf{t}) \cdot \Pr(\mathbf{t})$ in Eq. 3. Now, approximating the sum by the maximum, the optimization problem can be presented as

$$(\hat{\mathbf{t}},\hat{\mathbf{s}}) = \operatorname*{argmax}_{\mathbf{t},\mathbf{s}} \left(\Pr(\mathbf{s} \mid \mathbf{t}) \cdot \Pr(\mathbf{t}) \cdot \Pr(\mathbf{x} \mid \mathbf{s})\right), \tag{15}$$

and now the two-step approximation reduces to

$$\begin{aligned} \hat{\mathbf{s}} &\approx \operatorname*{argmax}_{\mathbf{s}} \{\Pr(\mathbf{s} \mid \mathbf{t}) \cdot \Pr(\mathbf{x} \mid \mathbf{s})\}, & (16)\\ \hat{\mathbf{t}} &\approx \operatorname*{argmax}_{\mathbf{t}} \Pr(\hat{\mathbf{s}} \mid \mathbf{t}) \cdot \Pr(\mathbf{t}) & (17)\\ &= \operatorname*{argmax}_{\mathbf{t}} \Pr(\hat{\mathbf{s}},\mathbf{t}). & (18) \end{aligned}$$

The main problem of this approach is the term $\mathbf{t}$ that appears in the first maximisation (Eq. 16). A possible solution is to follow an iterative procedure where $\mathbf{t}$, that is used for computing $\hat{\mathbf{s}}$, is the one obtained from $\operatorname{argmax}_{\mathbf{t}} \Pr(\hat{\mathbf{s}},\mathbf{t})$ in the previous iteration (García-Varea et al., 2000). In this case, $\Pr(\mathbf{s} \mid \mathbf{t})$ can be modelled by a source language model that depends on a previously computed $\tilde{\mathbf{t}}$: $\Pr_{\mathcal{N},\tilde{\mathbf{t}}}(\mathbf{s})$. In the first iteration no $\hat{\mathbf{t}}$ is known, but $\Pr_{\mathcal{N},\tilde{\mathbf{t}}}(\mathbf{s})$ can be approximated by $\Pr_{\mathcal{N}}(\mathbf{s})$. Following this idea, the search can be formulated as:

Initialization:

Let $\Pr_{\mathcal{N},\mathbf{t}}(\mathbf{s})$ be approximated by a source language model $\Pr_{\mathcal{N}}(\mathbf{s})$.

```
while not convergence
    1. Word decoding of $\mathbf{x}$. A source-language sentence $\hat{\mathbf{s}}$ is searched for using a source language model that depends on the target sentence, $\mathrm{Pr}_{\mathcal{N},\breve{\mathbf{t}}}(\mathbf{s})$, for $\mathrm{Pr}(\mathbf{s} \mid \mathbf{t})$ ($\breve{\mathbf{t}}$ is the $\hat{\mathbf{t}}$ computed in the previous iteration) and the corresponding HMMs, $\mathrm{Pr}_{\mathcal{M}}(\mathbf{x} \mid \mathbf{s})$, to model $\mathrm{Pr}(\mathbf{x} \mid \mathbf{s})$:
           $\hat{\mathbf{s}} \approx \operatorname*{argmax}_{\mathbf{s}} \left( \mathrm{Pr}_{\mathcal{N},\breve{\mathbf{t}}}(\mathbf{s}) \cdot \mathrm{Pr}_{\mathcal{M}}(\mathbf{x} \mid \mathbf{s}) \right).$
    2. Translation of $\hat{\mathbf{s}}$. A target-language sentence $\hat{\mathbf{t}}$ is searched for using a SFST, $\mathrm{Pr}_{\mathcal{T}}(\hat{\mathbf{s}}, \mathbf{t})$, as a model of $\mathrm{Pr}(\hat{\mathbf{s}}, \mathbf{t})$
           $\hat{\mathbf{t}} \approx \operatorname*{argmax}_{\mathbf{t}} \mathrm{Pr}_{\mathcal{T}}(\hat{\mathbf{s}}, \mathbf{t}).$
end of while
```

The first iteration corresponds to the sequential architecture proposed above.

While this seems a promising idea, only very preliminary experiments were carried out (García-Varea et al., 2000) and it has not been considered in the experiments presented in the present paper.

3 Experiments and results

Three sets of speech-to-speech translation prototypes have been implemented for Spanish to English and for Italian to English. In all of them, the application was the translation of queries, requests and complaints made by telephone to the front desk of a hotel. Three tasks of different degree of difficulty have been considered.

In the first one (EuTrans-0), Spanish-to-English translation systems were learned from a big and well controlled training corpus: about 170k different pairs ($\approx$ 2M running words), with a lexicon of about 700 words. In the second one (EuTrans-I), also from Spanish to English, the systems were learned from a random subset of 10k pairs ($\approx$ 100k running words) from the previous corpus; this was established as a more realistic training corpus for the kind of application considered. In the third and most difficult one, from Italian to English (EuTrans-II), the systems were learned from a small training corpus that was obtained from a transcription of a spontaneous speech corpus: about 3k pairs ($\approx$ 60k running words), with a lexicon of about 2,500 words.

For the serial architecture, the speech decoding was performed in a conventional way, using the same acoustic models as with the integrated architecture and trigrams of the source language models. For the integrated architecture, the speech decoding of an utterance is a sub-product of the translation process (the sequence of source words associated to the optimal sequence of transitions that produces the sequence of target words).

The acoustic models of phone units were trained with the HTK Toolkit (Woodland, 1997). For the EuTrans-0 and EuTrans-I prototypes, a training speech corpus of 57,000 Spanish running words was used, while the EuTrans-II Italian acoustic models were trained from another corpus of 52,000 running words

Performance was assessed on the base of 336 Spanish sentences in the case of EuTrans-0 and EuTrans-I and 278 Italian sentences in EuTrans-II. In all the cases, the test sentences (as well as the corresponding speakers) were different from those appearing in the training data.

For the easiest task, EuTrans-0, (well controlled and a large training set), the best result was achieved with an integrated architecture and a SFST obtained with the OMEGA learning technique. A Translation Word Error Rate of 7.6% was achieved, while the corresponding source-language speech decoding Word Error Rate was 8.4%. Although these figures may seem strange (and they would certainly be in the case of a serial architecture), they are in fact consistent with the fact that, in this task (corpus), the target language exhibits a significantly lower perplexity than the source language.

For the second, less easy task EuTrans-I, (well controlled task but a small training set), the best result was achieved with an integrated architecture and a SFST obtained with the MGTI learning technique (10.5% of word error rate corresponding to the speech decoding and 12.6% of translation word error rate).

For the most difficult task, EuTrans-II (spontaneous task and a small training set), the best result was achieved with a serial architecture and a SFST obtained with the MGTI learning technique (22.1% of word error rate corresponding to the speech decoding and 37.9% of translation word error rate).

4 Conclusions

Several systems have been implemented for speech-to-speech translation based on SFSTs. Some of them were implemented for translation from Italian to English and the others for translation from Spanish to English. All of them support all kinds of finite-state translation models and run on low-cost hardware. They are currently accessible through standard telephone lines with response times close to or better than real time.

From the results presented, it appears that the integrated architecture allows for the achievement of better results than the results achieved with a serial architecture when enough training data is available to train the SFST. However, when the training data is insufficient, the results obtained by the serial architecture were better than the results obtained by the integrated architecture. This effect is possible because the source language models for the experiments with the serial architecture were smoothed trigrams. In the case of sufficient training data, the source language model associated to a SFST learnt by the MGTI or OMEGA is better than trigrams (Section 2.1). However, in the other case (not sufficient training data) these source languages were worse than trigrams. Consequently an important degradation is produced in the implicit decoding of the input utterance.

Acknowledgments

The authors would like to thank the researchers that participated in the EuTrans project and have developed the methodologies that are presented in this paper.

This work has been partially supported by the European Union under grant IT-LTR-OS-30268, by the project TT2 in the “IST, V Framework Programme”, and Spanish project TIC 2000-1599-C02-01.

References

S. Bangalore and G. Ricardi. 2000. Stochastic finite-state models for spoken language machine translation. In *Workshop on Embeded Machine Translation Systems*.

S. Bangalore and G. Ricardi. 2001. A finite-state approach to machine translation. In *The Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

F. Casacuberta. 2000. Inference of finite-state transducers by using regular grammars and morphisms. In *Grammatical Inference: Algorithms and Applications*, volume 1891 of *Lecture Notes in Artificial Intelligence*, pages 1–14. Springer-Verlag.

I. García-Varea, A. Sanchis, and F. Casacuberta. 2000. A new approach to speech-input statistical translation. In *Proceedings of the International Conference on Pattern Recognition (ICPR2000)*, volume 2, pages 907–910, Barcelona, Sept. IAPR, IEEE Press.

D. Llorens. 2000. *Suavizado de autómatas y traductores finitos estocásticos*. Ph.D. thesis, Universitat Politècnica de València.

H. Ney, S. Nießen, F. Och, H. Sawaf, C. Tillmann, and S. Vogel. 2000. Algorithms for statistical translation of spoken language. *IEEE Transactions on Speech and Audio Processing*, 8(1):24–36.

H. Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proceedins of the IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 517–520, Phoenix, AR, March.

D. Picó and F. Casacuberta. 2001. Some statistical-estimation methods for stochastic finite-state transducers. *Machine Learning*, 44:121–141.

J.M. Vilar. 2000. Improve the learning of subsequential transducers by using alignments and dictionaries. In *Grammatical Inference: Algorithms and Applications*, volume 1891 of *Lenture Notes in Artificial Intelligence*, pages 298–312. Springer-Verlag.

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