FINITE-STATE TRANSDUCERS FOR SPEECH-INPUT TRANSLATION

F. Casacuberta

Dpt. de Sistemes Informàtics i Computació, Institut Tecnològic d'Informàtica
Universitat Politècnica de València, 46071 València, SPAIN.
fcn@iti.upv.es

ABSTRACT

Nowadays, hidden Markov models (HMMs) and n-grams are the basic components of the most successful speech recognition systems. In such systems, HMMs (the acoustic models) are integrated into a n-gram or a stochastic finite-state grammar (the language model). Similar models can be used for speech translation, and HMMs (the acoustic models) can be integrated into a finite-state transducer (the translation model). Moreover, the translation process can be performed by searching for an optimal path of states in the integrated network. The output of this search process is a target word sequence associated to the optimal path. In speech translation, HMMs can be trained from a source speech corpus, and the translation model can be learned automatically from a parallel training corpus.

This approach has been assessed in the framework of the EuTrans project, founded by the European Union. Extensive speech-input experiments have been carried out with translations from Spanish to English and from Italian to English translation, in an application involving the interaction (by telephone) of a customer with a receptionist at the front-desk of a hotel. A summary of the most relevant results are presented in this paper.

1. INTRODUCTION

The statistical framework has proved very useful in automatic speech recognition. This paradigm is based on automatically built acoustic models from sufficiently large speech training sets and the language model from sufficiently large text training set. An interesting feature of this approach is the type of search used to obtain a decoded sentence. This search is performed on an integrated network of acoustic models in a language model. This integration can be carried out through the use of stochastic finite-state networks as models, particularly hidden Markov models as acoustic models and n-grams or stochastic finite-state grammars as language models.

This work has been partially supported by the European Union under grant IT-LTR-OS-30268.

This approach can be naturally extended to the development of (speech-input) machine translation systems. On the one hand, the acoustic models can be trained in a way similar to that used in speech recognition. On the other hand, the translation model can be learned from a sufficiently large training set of parallel-text, using adequate learning algorithms [1, 2].

The possibility of using stochastic finite-state transducers for limited-domain translation has been discussed in previous works [3, 4, 5]. These models obviously support the integrated architecture as well as the serial architecture. The later architecture is usually adopted in the speech translation prototypes proposed so far. In the integrated architecture, the acoustic models are *integrated* in the translation model in a way similar to that used in speech recognition. The search procedure for translation is based on the very same *Viterbi* search engine used in speech recognition. A similar approach has been proposed in [6, 7].

In this paper, we present the main results that have been achieved for speech translation using finite-state methodologies. These methodologies were developed in the EuTrans project. EuTrans was a five-year joint effort of four European institutions (http://www.zeres.de/ eutrans) which was partially funded by the *Open Domain* of the *Long-Term Research (LTR)* ESPRIT program of the European Union.

2. FINITE-STATE TRANSDUCERS AND MACHINE TRANSLATION

2.1. The statistical framework for machine translation

Let **s** be a source sentence. The translation of **s** into a target language can be formulated as the search for a word sequence, $\hat{\mathbf{t}}$, from a target language such that:

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

$\Pr(\mathbf{s}, \mathbf{t})$ can be approximated using a *stochastic finite-state transducer* (SFST) as a *translation model* [8]. Other different types of models have been proposed elsewhere [9, 10, 11].

2.2. Finite-state transducers

A SFST, $\mathcal{T}$, is a tuple $<Q, \Sigma, \Delta, R, q_0, F, P>$, where:

(a) Q is a finite set of *states*

(b) q_0 is the *initial state*

(c) Σ is a finite set of *input symbols* (source words)

(d) Δ is a finite set of *output symbols* (target words) ($\Sigma \cap \Delta = \emptyset$)

(e) 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]

(f) $P : R \rightarrow I\!R^+$ (*transition probabilities*) and $F : Q \rightarrow I\!R^+$ (*final-state probabilities*) are functions that $\forall q \in Q$:

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

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

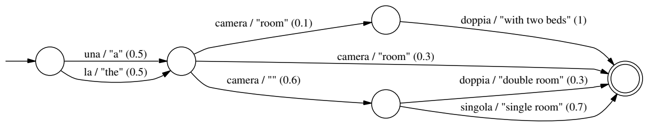


Fig. 1. Example of a SFST. "" denotes the empty string. The source sentence *"una camera doppia"* can be translated to either *"a double room"* or *"a room with two beds"*. The most probable translation is the first one with probability of 0.09.

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 $(q, a, \omega, q), (q, a, \omega', q') \in R$, then $\omega = \omega'$ and $q = q'$) [2].

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

$$d(\mathbf{s}, \mathbf{t}) : (q_0, \mathrm{s}_1, \tilde{\mathrm{t}}_1, q_1), (q_1, \mathrm{s}_2, \tilde{\mathrm{t}}_2, q_2), \ldots, (q_{I-1}, \mathrm{s}_I, \tilde{\mathrm{t}}_I, q_I),$$

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 a translation form, $d(\mathbf{s}, \mathbf{t})$, is:

$$\Pr_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})) = \prod_{i=0}^{I} P(q_{i-1}, \mathrm{s}_i, \tilde{\mathrm{t}}_i, q_i) \cdot F(q_I). \qquad (2)$$

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

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

$$\Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \sum_{d(\mathbf{s},\mathbf{t})} \Pr_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})). \qquad (3)$$

Using $\Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t})$ as an approximation to $\Pr(\mathbf{s}, \mathbf{t})$ in Eq. 1, the stochastic translation of a source sentence $\mathbf{s} \in \Sigma^\star$ by a SFST $\mathcal{T}$ is given by

$$\underset{\mathbf{t}}{\text{argmax}} \Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t}). \qquad (4)$$

The probability of Eq. 3 can be approximated by using the maximisation over all possible translation forms instead of the sum,

$$\Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) \approx \max_{d(\mathbf{s},\mathbf{t})} \Pr_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})). \qquad (5)$$

In this case, the stochastic translation of $\mathbf{s}$ by a SFST $\mathcal{T}$ can be approximately computed using the Viterbi algorithm. This algorithm can search for the optimal sequence of states in the SFST $\mathcal{T}$ that deals with $\mathbf{s}$. The translation of $\mathbf{s}$ is the concatenation of target strings that are associated to the optimal sequence of transitions [12].

These models have implicit source and target language models in their definitions. In practice, the source language model can be obtained by removing the target words from each transition. On the other hand, the target language model can be obtained by removing the source words from each transition.

2.3. Learning finite-state transducers

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 *Morphic Generator Translator Inference (MGTI)* technique [1]. Alternatively, the structural component can be learned using the *OSTIA Modified for Employing Guarantees and Alignments (OMEGA)* technique [2]. In this case, the probabilistic components can be estimated in a second step using a *maximum likelihood* technique or other possible criteria [12]. One of the main problems that appears during the learning process is the modelling of the events that have not been seen in the training set. This problem can be confronted in a way similar to the one used in language modelling by using smoothing techniques in the estimation process of the probabilistic components of the SFST [13]. Alternatively, the smoothing can be considered in the process of learning both components [1].

The MGTI is based on the following idea for the inference of a transducer: given a finite sample of string pairs [1]:

1. **Building training strings.** Each training pair is transformed into a single string from an *extended alphabet* to obtain a new sample of strings.

2. **Inferring a (stochastic) regular grammar.** Typically an N-gram is inferred from the sample of strings obtained in the previous step.

3. **Transforming the inferred regular grammar into a transducer.** The symbols associated to the grammar rules are transformed into input/output symbols by applying an adequate transformation, thereby transforming the grammar inferred in the previous step into a transducer.

The transformation of a parallel corpus into a string corpus is performed using statistical alignments (a function from the set of positions in the target sentence to the set of positions in the source sentence) [9, 10, 11]. A training string is built by assigning the corresponding aligned word from source sentence to each word from the target sentence. This assignment must not violate the order in the target sentence [1]. Using this type of transformation from a pair of strings into a string of extended symbols, the transformation from a grammar to a finite-state transducer in step 3 is straightforward.

An interesting feature of the MGTI method is that all the techniques which are known for n-gram smoothing are readily applicable in the second step of the method.

The OMEGA algorithm [2] can be seen as an improvement over OSTIA [14], a previous algorithm for inferring SSTs.

There are two main phases in the training:

1. **Building an initial tree** for representing the samples, where the prefixes of the source sentences are represented in a compact mode and the target sentences are in the leaves of the tree.

2. **State merging**. A series of merges of states is carried out in order to generalize the training corpus. After each merge of states, the resulting graph should deal with the training set and possibly other pairs.

The sequence of merges is performed, in the initial tree, level by level taking into account the states in the previous levels.

Source and/or target language models are used to determine whether two states can be merged. Another criterium for joining states makes use of statistical dictionaries and alignments [9, 10, 11].

3. FINITE-STATE TRANSDUCERS AND SPEECH TRANSLATION

3.1. The statistical framework for speech translation

Let $\mathbf{x}$ be an acoustic representation of a given utterance. The translation of $\mathbf{x}$ into a target language can be formulated as the search for a word sequence, $\hat{\mathbf{t}}$, from a target language such that:

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t} \mid \mathbf{x}). \tag{6}$$

Conceptually, the translation can be viewed as a two-step process [15, 11]:

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

where $\mathbf{s}$ is a possible decoded sequence of $\mathbf{x}$ in the source language whose word sequence in the target language is $\mathbf{t}$. Consequently,

$$\operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}|\mathbf{x}) = \operatorname*{argmax}_{\mathbf{t}} \sum_{\mathbf{s}} \Pr(\mathbf{t}, \mathbf{s}|\mathbf{x}), \tag{7}$$

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

$$\operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}|\mathbf{x}) = \operatorname*{argmax}_{\mathbf{t}} \left(\sum_{\mathbf{s}} \Pr(\mathbf{s}, \mathbf{t}) \cdot \Pr(\mathbf{x}|\mathbf{s}) \right). \tag{8}$$

In practice, the sum in Eq. 8 can be approximated with a maximisation. This simplification allows the simultaneous computations of the decoded sentence and the target sentence:

$$\operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}|\mathbf{x}) \approx \operatorname*{argmax}_{\mathbf{t}} \max_{\mathbf{s}} \left(\Pr(\mathbf{s}, \mathbf{t}) \cdot \Pr(\mathbf{x}|\mathbf{s})\right). \tag{9}$$

$\Pr(\mathbf{x}|\mathbf{s})$ is usually approximated using *acoustic models*, typically hidden Markov models [16] and $\Pr(\mathbf{s}, \mathbf{t})$ is approximated using a *translation model*. As for machine translation (Section 2), SFSTs are models that allow a direct approach to this probabilistic distribution $\Pr(\mathbf{s}, \mathbf{t})$.

3.2. Architectures for speech translation

Using Eq. 3 (or Eq. 5) as an approach to $\Pr(\mathbf{t}, \mathbf{s})$, and HMMs as approaches to $\Pr(\mathbf{x}|\mathbf{s})$, Eq. 9 is transformed in the optimization problem:

$$\max_{\mathbf{s},\mathbf{t}} \left(\Pr_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) \cdot \Pr_{\mathcal{M}}(\mathbf{x}|\mathbf{s})\right), \tag{10}$$

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

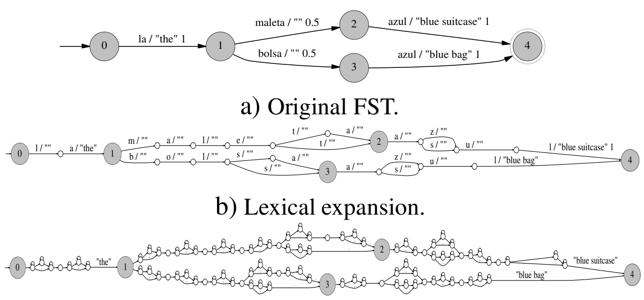
The computation of the most likely target sentence $\mathbf{t}$ for an observed *acoustic source sentence* $\mathbf{x}$ is accomplished using a search algorithm for the optimization problem in Eq. 10.

For the sake of simplicity in this section, we assume that $\mathbf{x}$ is segmented in I acoustic subsequences and that each sequence is associated to one source word:

$$\Pr_{\mathcal{M}}(\mathbf{x} \mid \mathbf{s}) = \prod_{i=1}^{I} \Pr_{\mathcal{M}}(\bar{\mathbf{x}}_i \mid \mathbf{s}_i), \tag{11}$$

where $\bar{\mathbf{x}}_i$ is the i acoustic segment, and each source word $\mathbf{s}_i$ has a HMM associated that supplies the density values $\Pr_{\mathcal{M}}(\bar{\mathbf{x}}_i|\mathbf{s}_i)$. Therefore, if Eq. 5 is used as $\Pr_{\mathcal{T}}(\mathbf{s},\mathbf{t})$, Eq. 10 can be rewritten

$$\max_{\mathbf{s},\mathbf{t}} \max_{d(\mathbf{s},\mathbf{t})} \prod_{i=1}^{I} P(q_{i-1}, \mathbf{s}_i, \bar{\mathbf{t}}_i, q_i) \cdot \Pr_{\mathcal{M}}(\bar{\mathbf{x}}_i|\mathbf{s}_i). \qquad (12)$$



a) Original FST.

b) Lexical expansion.

c) Phonetic expansion.

Fig. 2. Example of the integration process of the lexical knowledge (figure b) and the phonetic knowledge (figure c) in a FST (figure a). “” denotes the empty string.

From Eq. 12, the search for an optimal $\mathbf{t}$ (and $\mathbf{s}$) can be viewed as the search problem for the optimal sequence of states in an integrated network (*integrated architecture*). This network is built by a process of substitution of the edges of the SFST by the corresponding HMM of the source word associated to the edge. This integration is shown in Fig. 2. A small SFST is presented in the part a) of this figure. In part b), the source words in each edge are substituted by the corresponding phonetic transcription. In part c), each phoneme is substituted by the corresponding HMM of the phone.

Sometimes, the integrated network can be very huge and the search for the optimal target can require a high computational effort. Heuristic techniques such as beam search can be used to reduce the computational cost of the search. An alternative way of reducing this computational effort can be achieved by a further approximation consisting in breaking the search down into two steps in a *“serial architecture”*. In this case, there is a conventional source speech decoding that is followed by a translation of the decoded sentence into the target sentence.

Using $\Pr(\mathbf{t},\mathbf{s}) = \Pr(\mathbf{t} \mid \mathbf{s}) \cdot \Pr(\mathbf{s})$ in Eq. 9, the optimization problem can be presented as

$$\operatorname*{argmax}_{\mathbf{t}} \max_{\mathbf{s}} \left\{ \Pr(\mathbf{t}|\mathbf{s}) \cdot \Pr(\mathbf{s}) \cdot \Pr(\mathbf{x}|\mathbf{s}) \right\}. \qquad (13)$$

The search for the optimal target sentence to Eq. 13 can be approximated as follows:

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

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

2. *Translation of* $\hat{\mathbf{s}}$. The target sentence $\hat{\mathbf{t}}$ is searched using a SFST, $\Pr_{\mathcal{T}}(\hat{\mathbf{s}},\mathbf{t})$, as an approximation to $\Pr(\hat{\mathbf{s}},\mathbf{t})$ ($\operatorname{argmax}_{\mathbf{t}} \Pr(\mathbf{t} \mid \hat{\mathbf{s}}) = \operatorname{argmax}_{\mathbf{t}} \Pr(\mathbf{t},\hat{\mathbf{s}})$):

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

4. EXPERIMENTS AND RESULTS

4.1. The ATROS system

The speech translation system used in the experiments was based on the ATROS (Automatically Trainable Recognizer Of Speech) engine [5]. 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. A first version of ATROS for Spanish continuous speech recognition was presented in [17].

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 a source sentence is built by concatenating the target strings of the successive transitions that compose the optimal path.

Obviously, the ATROS system also supports the serial architecture. In this case, a source language model is supplied to the ATROS system in order to compute a decoded source sentence. In a second step, ATROS can translate the decoded source sentence into the target sentence by using a SFST.

4.2. Tasks and corpora

Two translation tasks of different degrees of difficulty were used for the experiments. The acoustic data was acquired by telephone.

In the first one, the translation from Spanish to English (EuTrans-0), the SFSTs were learned with a controlled corpus of 490,000 pairs [4]. The size of the vocabulary was 686 Spanish words and 513 English words. The bigram test-set perplexities were 8.6 for Spanish and 5.2 for English. The acoustic models were 26 continuous density HMMs corresponding to a set of 26 Spanish phone units. The acoustic models were trained with the HTK Toolkit [18] using an Spanish corpus of 11,000 running words from 20 speakers. The speech test set was composed of 336 Spanish sentences uttered by four speakers.

In the second task (EuTrans) [5], the SFST was learned with a training corpus of 3,038 pairs that was obtained from a transcription of a spontaneous speech corpus. The size of the vocabulary was 2,459 Italian words and 1,701 English words. The bigram test-set perplexities were 31 for Italian and 25 for English. The acoustic models were context-dependent continuous density HMMs selected by the CART method and trained using the Viterbi approach [19]. The speech training set was composed of 52,511 running words. The speech test set was composed by 278 Italian sentences which had not been used in training.

4.3. Experimental results

The system assessment was performed using two error criteria. One of them was the *Word Error Rate (WER)* of the source decoded sentence. The second criterium was the *Translation Word Error Rate (TWER)* of the target sentence. Both values were computed by comparing the decoded sentence or the translated sentence, respectively, with the corresponding reference (source or target) sentences using fractional programming techniques.

The results presented with a serial architecture were achieved using a trigram language model for the input speech decoding.

The Italian-English EuTrans prototype achieved quite an acceptable response time (about three times real time or less), while the Spanish-English EuTrans-0 prototype often run in less than real time, even on low-cost Pentium machines.

The results achieved in the EuTrans-0 and EuTrans are presented in Table 1. A complete set of results will be available in [20].

For the easiest task EuTrans-0 (controlled task and a large training set), the results achieved with an integrated architecture were better than the results achieved with the serial architecture for both, MGTI and OMEGA learning techniques. However, for the most difficult task EuTrans (spontaneous task and a small training set), the results with an integrated architecture were worse than the results achieved with the serial architecture for both, MGTI and OMEGA learning techniques.

Table 1. Experimental results achieved with the integrated and the serial approaches. "Arch" stands for architecture: integrated (INT) or serial (SER). "SLM" is the source language model used in the experiment: the implicit model in the SFST learnt by OMEGA or MGTI (for the integrated architecture), or a 3-grams (for the serial architecture).

EuTrans-0				
Models	Arch	SLM	WER(%)	TWER(%)
OMEGA	INT	OMEGA	8.4	7.6
OMEGA	SER	3-grams	8.6	9.4
MGTI	INT	MGTI	7.5	10.7
MGTI	SER	3-grams	8.6	11.6

EuTrans				
Models	Arch	SLM	WER(%)	TWER(%)
MGTI	SER	3-grams	22.1	37.9
MGTI	INT	MGTI	32.0	44.8
OMEGA	SER	3-grams	22.1	49.4
OMEGA	INT	OMEGA	52.5	57.0

5. DISCUSSION AND 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 were implemented for translation from Spanish to English. All of them 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 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 2.2). 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. To overcome the problem of learning SFST for speech translation with small amounts of training data, it is necessary to improve the available learning techniques in order to produce SFST with good source language models.

Acknowledgments

The author would like to thank to the researchers that participated in the EuTrans project and have developed the methodologies that are presented in this paper. In particular, the author would like to mention Hermann Ney and Enrique Vidal, the leaders of two of the teams involved in the project, as well as Alberto Sanchis for the large number of experiments that were carried out.

6. REFERENCES

[1] 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.

[2] 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.

[3] E.Vidal, "Finite-state speech-to-speech translation," in *Proceeding of the IEEE International Conference on Acoustic Speech and Signal Processing*, 1997, pp. 111–114.

[4] 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*, vol. 15, no. 1-2, pp. 75–103, 2001.

[5] F. Casacuberta, D. Llorens, C. Martínez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Picó, A. Sanchis, E. Vidal, and J. M. Vilar, "Speech-to-speech translation based on finite-state transducers," in *Proceedins of the IEEE International Conference on Acoustic, Speech and Signal Processing*, 2001, pp. 613–616.

[6] S. Bangalore and G. Ricardi, "Stochastic finite-state models for spoken language machine translation," in *Workshop on Embeded Machine Translation Systems*, 2000.

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

[8] F.Casacuberta, "Maximum mutual information and conditional maximum likelihood estimation of stochastic regular syntax-directed translation schemes," in *Grammatical Inference: Learning Syntax from Sentences*, vol. 1147 of *Lecture Notes in Artificial Intelligence*, pp. 282–291. Springer-Verlag, 1996.

[9] P.F.Brown, J.Cocke, S.A.Della Pietra, V.J.DellaPietra, F.Jelinek, J.D.Lafferty, R.L.Mercer, and P.S.Roosin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[10] P.F.Brown, S.A.Della Pietra, V.J.Della Pietra, and R.L.Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–310, 1993.

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

[12] D. Picó and F. Casacuberta, "Some statistical-estimation methods for stochastic finite-state transducers," *Machine Learning*, vol. 44, pp. 121–141, 2001.

[13] D. Llorens, *Suavizado de autmatas y traductores finitos estocsticos*, Ph.D. thesis, Universitat Politècnica de València, 2000.

[14] E.Vidal J.Oncina, P.Garcia, "Learning subsequential transducers for pattern recognition interpretation tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 5, pp. 448–458, 1993.

[15] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proceedins of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Phoenix, AR, Mar. 1999, pp. 517–520.

[16] K. Knill and S. Young, *Corpus-Based Statiscal Methods in Speech and Language Processing. eds.: S. Young and G.Bloothooft*, chapter Hidden Markov Models in Speech and Language Processing, pp. 27–68, Kluwer Academic Publishers, 1997.

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

[18] 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.

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

[20] F. Casacuberta, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, H. Ney, F. Och, M. Pastor, D. Picó, A. Sanchis, E. Vidal, and J. M. Vilar, "Statistical and finite-state approaches to speech-to-speech translation," To be published.