TORNASOL: AN INTEGRATED SYSTEM FOR THE CONTINUOUS SPEECH RECOGNITION OF SPANISH

L.J. Rodríguez, A. Varona, K. López de Ipiña and M.I. Torres

Departamento de Electricidad y Electrónica. Facultad de Ciencias.
Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU)
Apartado 644. 48080 Bilbao. Spain.
e-mail: luisja@we.lc.ehu.es

Abstract

This paper presents a new system for the continuous speech recognition of Spanish, integrating previous works in the fields of acoustic-phonetic decoding and language modelling. The system includes decision tree-based sublexical units and syntactic language models based on regular grammars. Acoustic and language models -separately trained with speech and text samples, respectively- are integrated into a single stochastic finite state automaton. Acoustic and language model probabilities were heuristically balanced and then combined according to a standard beam search procedure. The system was evaluated over a task oriented Spanish speech corpus consisting of 82.000 words and a vocabulary of 1.213 words, resulting in more than 90% word recognition rate.

Keywords: Continuous Speech Recognition, Decision Trees, Stochastic Grammars

1. Introduction

This paper describes *Tornasol*[1]: a Continuous Speech Recognition (CSR) System developed at the University of the Basque Country for medium/large size Spanish tasks. The system includes, as its main features:

- Context dependent sublexical units obtained by applying Decision Tree based clustering to a phonetically balanced Spanish speech database.
- Syntactic language models based on regular grammars, fully integrated in the recognition system.
- A linear organization of the lexicon fully compatible with the search strategy.
- A client/server architecture.

[1] The name *Tornasol* was chosen in honour of the famous deaf scientist *Silvestre Tornasol* (*Professor Calculus* in the english version) appearing in the comic series *Tintin*.

Some state-of-the-art speech recognition technologies are also included: robust speech signal analysis, discrete hidden Markov models, beam-search and (optionally) fast phoneme look ahead algorithms.

The CSR System was evaluated over a task oriented speech corpus representing a set of queries to a Spanish geography database. This is a specific task designed to test integrated systems (involving acoustic, syntactic and semantic models) in automatic speech understanding.

The paper is organized as follows: Section 2 summarizes the speech analysis procedure and the acoustic models. In Section 3 the language models and the search strategy are briefly outlined. Section 4 presents the evaluation of the system: firstly reference and target sets of sublexical units are evaluated over an acoustic-decoding task; then the whole CSR system is evaluated over a Spanish recognition task. Section 5 concludes the paper and notes outstanding issues for future research.

2. Acoustic-phonetic modelling

Speech analysis -as suggested by [1]- is made by following conventions of the Entropic Hidden Markov Modeling Toolkit (HTK) [2], with some minor changes. Speech -acquired typically through a headset condenser microphone- is sampled at 16 kHz, each sample having 16 bit precision. Speech samples are grouped into successive frames and passed through a Hamming window. Frame length is 25 ms and inter-frame distance 10 ms. A filter bank formed by 24 triangular filters with increasing widths according to mel-frequency scale, is applied to a 512-point FFT, producing 24 spectral weighted mean values. A discrete cosine transform applied to these coefficients decorrelates the spectral envelope from other characteristics not relevant for recognition, producing 12 mel-frequency cepstral coefficients (MFCC). Cepstral mean normalization and liftering are then applied to compensate for channel distortion and speaker characteristics. The normalized logarithm of the energy is also computed. Finally, dynamic characteristics (first and second derivatives) are added to the MFCC and log-energy parameters, obtaining a 39-component acoustic observation vector.

As usual when dealing with medium to large size vocabularies, sublexical units were used as the basic speech units. Three different sets were defined. The first and simpler one contained 24 phone-like units, including a single model for silence. The second one was a mixture of 103 trainable (e.g. with enough training samples) monophones, diphones and triphones, as described in [3]. The well known technique of decision tree clustering [4] was applied to obtain a third optimal set of 101 trainable, discriminative and generalized context dependent units. An additional set of border units was specifically trained to generate lexical baseforms, covering all possible intraword contexts and being context independent to the outside. We are currently working in generating a more general set of inter and intraword context dependent units, and a first approach was made to evaluate their contribution to the recognition process (see Section 4). To deal with multiple

codebook observations, as was the case in our baseline system, a simple and not very expensive discriminative function, combining probabilities from all the codebooks, was used. The set of decision tree based context dependent units gave the best results in acoustic-phonetic decoding experiments and also when building word models for a more reliable test, as will be shown in Section 4.

Discrete Left-to-Right Hidden Markov Models with 3 looped states and 4 discrete observation distributions per state, were used as acoustic models. Emision and transition probabilities were estimated using both the Baum-Welch and Viterbi procedures, applying the Maximum Likelihood criterion.

To deal with discrete acoustic models, the standard LBG Vector Quantization procedure was applied, obtaining four different codebooks, each containing 256 centroids, corresponding to MFCCs, first derivatives of MFCCs, second derivatives of MFCCs and a 3-component vector formed by log-energy, first and second derivatives of log-energy. During recognition, each parameter subvector was assigned the nearest centroid index and a four index tag was passed as acoustic observation to the search automaton.

3. Language modelling and search strategy

A syntactic approach to the well known *n-gram* models, the k-Testable in the Strict Sense (k-TSS) language models, which can be viewed as a kind of stochastic regular grammars -thus defining a subclass of regular languages-, was used. The use of k-TSS language models allowed the construction of a deterministic, and hence unambiguous, Stochastic Finite State Automaton (SFSA) integrating a number of k-TSS language models in a self-contained model [5]. Then a syntactic back-off smoothing technique was applied to the SFSA to consider unseen events [6]. This formulation strongly reduced the number of parameters to be handled and led to a very compact representation of the model parameters learned at training time. Thus the smoothed SFSA was efficiently allocated in a simple array of an adequate size [7].

Lexical baseforms were linearly represented. Each word was transcribed as the concatenation of sublexical units corresponding to its standard pronunciation, taking into account only intraword contexts. Thus word nodes were expanded with the corresponding concatenation of previously trained acoustic models, and one single automaton resulted including both acoustic and syntactic probabilities.

The time-synchronous Viterbi decoding algorithm was used at parsing time: a simple search function through the array representing the Language Model allowed a straightforward access to the next state of the SFSA, needed at each decoding time. After some preliminary experimentation, a kind of balance between acoustic and syntactic probabilities was found necessary in the search automaton. A weight α affecting the Language Model probabilities was heuristically optimized. Then a beam threshold was established and (optionally) a fast phoneme look ahead algorithm was included to reduce the average size of the search network [5].

4. Experimental evaluation

All the parts of the system were carefully tested and optimized before integrating them. From the audio acquisition module to the search module, separate evaluations were made. Here we present only the selection of an adequate set of sublexical units for acoustic modelling, and the integration of acoustic and language models into one single search automaton, which constitutes the core of the CSR system.

4.1. Evaluating sublexical units

As mentioned above, three different sets of sublexical units were tested: 24 context independent phone-like units, a mixture of 103 monophones, diphones and triphones, and 101 decision tree based generalized context dependent units. An acoustic-phonetic decoding task was used as benchmark, having a balanced phonetic corpus both for training and testing purposes. The training corpus was composed of 1529 sentences, involving around 60000 phones. The -speaker independent- test corpus was composed of 700 sentences. Note that we did not apply any phonological model in these experiments. Table 1 shows the results.

Table 1. Phone recognition rates for a Spanish acoustic-phonetic decoding task, using three different sets of sublexical units, without any phonological model.

Type of unit	# units	% REC
CI phone-like	24	63.97
Mixture (Mph, Dph, Tph)	103	65.90
DT-based	101	66.44

Context independent phone-like units gave significative lower rates (around two points) than the two other sets of context dependent units, which in turn showed similar performances.

However, a more reliable test seems necessary to decide which set of units is more suitable, by building lexical baseforms and obtaining word recognition rates. Two different procedures were used to obtain lexical baseforms, both considering words as isolated. In the first one, called TR1, border units were selected to be context independent both sides. In the second one, called TR2, border units were selected to be context dependent in the word side and context independent in the outside. Another transcription procedure, called TR3, was used to test the contribution of modelling interword contexts to speech recognition. TR3 takes into account the interword contexts appearing in the test corpus to obtain lexical baseforms of words. Table 2 shows word recognition rates using these transcription procedures over the same test corpus mentioned above. Only the baseforms of the 203 words found in the test corpus were used to run the alignment procedure, without applying any language model.

Table 2. Word recognition rates for the same test corpus used in Table 1, and three different transcription procedures. Language Model probabilities were not applied.

	% Word Recognition		
Type of unit	**TR1**	**TR2**	**TR3**
CI phone-like	49.83	-	
Mixture (Mph, Dph, Tph)	-	51.16	56.73
DT-based	52.86	53.26	58.01

Word recognition rates show that TR2 works better than TR1 -as may be expected. However, TR3 gave the best performance, showing the importance of modelling interword contexts. Finally, note that DT-based context dependent units outperformed the two other sets of units.

4.2. Evaluating the whole system

The system is designed as a distributed client-server application. The server is the main program, as it includes the search module. On the other hand, the client application acquires the audio signal, does the preprocessing and the feature extraction, and includes the graphical user interface [8].

In the search procedure, not only the parameter α weighting language model probabilities over acoustic probabilities must be optimized, but also the beam parameter. Actually, both should be optimized jointly, but we first heuristically found an optimum a and then an optimum beam. The beam is a key issue to reach real-time operation, so a balance between system accuracy and computation time must be found.

At any time, each possible transition from each active node in the search network has an acoustic probability and possibly also a syntactic probability, which are multiplied by the accumulated probability at the departure node, and assigned as accumulated probability to the arrival node. To obtain the set of active nodes at that time, the beam parameter must be applied to discard all nodes whose accumulated probability falls below certain threshold, thus drastically reducing the size of the search network. This threshold is usually obtained by multiplying the beam parameter by the maximum accumulated probability at that time.

Both context independent phone-like and DT-based sublexical units were used in these experiments. Lexical baseforms were generated by applying TR1 and TR2 to define border units. A task-oriented Spanish speech corpus [9], consisting of 82.000 words and a vocabulary of 1.213 words, was used. This corpus represents a set of queries to a Spanish geography database. The training corpus for the *k*-TSS language models consisted of 9150 sentences. For testing purposes, an independent but fully covered text containing 200 sentences was used. These sentences were uttered by 12 speakers resulting in a total of 600 sentences and 5655 words. Table 3 shows the system performance for k=2, 3 and

4, beam=0.67, 0.55 and 0.45, and fixed α=6, using only context independent phone-like units.

Table 3. System performance for different values of k and beam, and fixed α=6. Significative measures are showed: average number of active nodes (#AN average), average processing time per frame (PTF average, in milliseconds), word recognition rate (%W) and sentence recognition rate (%S). Context independent phone-like units were used as sublexical units.

k	beam	#AN average	PTF average (ms)	%W	%S
	0.67	40.13	3.80	68.58	29.17
2	0.55	218.21	11.75	84.05	41.67
	0.45	858.51	33.42	85.19	44.17
	0.67	32.24	3.30	70.48	36.33
3	0.55	179.01	10.96	89.15	56.00
	0.45	800.52	36.67	90.35	57.50
	0.67	31.50	3.38	71.36	45.00
4	0.55	177.99	11.24	89.78	59.00
	0.45	808.30	38.19	91.42	61.50

It can be seen that computational resources required by the system (processing time and memory, i. e. the number of active nodes) reduce drastically as the beam narrows, while system accuracy remains quite good, especially for k=4. Note that only the more constrained search (with beam=0.67) fullfils the condition of real-time operation (i.e. average processing time per frame less than 10 milliseconds). Note also that the intermediate beam (0.55) is very close to that condition.

Table 4 shows the system performance for k=2, 3 and 4, fixed α=6 and fixed beam=0.55, with both context independent phone-like and DT-based sublexical units, using transcription procedures TR1 and TR2 for the latter.

Table 4. System performance for different values of k, fixed beam=0.55 and α=6. Significative measures are showed: average number of active nodes (#AN average), word recognition rate (%W) and sentence recognition rate (%S). CI phone-like units and DT-based sublexical units (TR1 and TR2) are compared.

	CI phone-like			DT-based (TR1)			DT-based (TR2)		
k	#AN average	%W	%S	#AN average	%W	%S	#AN average	%W	%S
2	218	84.05	41.67	173	85.25	42.33	134	85.72	46.67
3	179	89.15	56.00	143	89.31	56.17	108	89.09	59.50
4	178	89.78	59.00	142	89.80	59.17	107	89.31	60.50

This time DT-based sublexical units only outperformed context independent phone-like units for k=2. This could discourage us from using the Decision Tree based

clustering approach to define a suitable set of sublexical units. However, the use of DT-based sublexical units always reduced the average number of active nodes in the search network, and thus the time required to decode each frame. This effect was stronger when the transcription procedure TR2 was applied to define the border units. In fact, the more suitable result -marked on the table- was obtained for k=4, with DT-based sublexical units and the transcription procedure TR2. The word recognition rate (89.31) was only slightly lower than those obtained with CI phone like units (89.78) and DT-based units with TR1 (89.80), giving the highest sentence recognition rate (65.50) with the lowest memory requirements (107 active nodes average). So this would be a good choice in order to fulfill the condition of real time operation with a good system accuracy.

5. Concluding remarks

In this paper a new CSR system -called *Tornasol*- for medium/large size Spanish tasks was presented. State-of-the-art speech recognition technologies, including robust feature extraction, discrete HMMs and a time-synchronous Viterbi decoding algorithm provided with beam search and (optionally) fast phoneme look ahead, were integrated into a client/server architecture, where the main search module was implemented as a server, and the acoustic front-end was part of the client application.

However, the most remarkable contributions were the use of a Decision Tree based clustering algorithm to define a suitable set of sublexical units -which increased the discrimination among acoustic models and allowed a more flexible and generalized way of building lexical baseforms-, and the use of a kind of stochastic regular grammars, the k-TSS language models, which led to significative reductions in computational requirements.

Future work includes expanding the currently available client, designed for a Silicon Graphics workstation, to more common platforms, like PCs with UNIX; the use of Decision Tree clustering to form lexical baseforms incorporating not only intraword but also interword context modelling; the use of more accurate acoustic models, and finally improving the search module to fully acomplish real-time operation with larger vocabulary sizes, while keeping good system accuracy.

Acknowledgements

This work is part of two greater CSR projects supported by the Spanish CICYT (codes TIC95-0884-C04-03 and TIC98-0423-C06), jointly developed with four other Spanish universities: Universidad Politécnica de Valencia, Universidad Politécnica de Cataluña, Universidad de Zaragoza and Universitat Jaume I. Here we would like to thank all the help and advise obtained from them.

References

[1] J.B. Mariño, J. Hernando. “Notes on feature extraction for speech recognition”. *Personal communication. March 1998.*

[2] S. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland. “The HTK Book. Hidden Markov Model Toolkit V. 2.1”. *Entropic Cambridge Research Laboratory. March 1997.*

[3] A. Bonafonte, R. Estany, E. Vives. “Study of subword units for Spanish speech recognition”. *Proc. EUROSPEECH-95*, pp.1607-1610.

[4] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny. “Decision Trees for Phonological Rules in Continuous Speech”. *Proc. IEEE ICASSP-91*, pp. 185-188.

[5] A. Varona, I. Torres. “Using Smoothed *k*-TSS Language Models in Continuous Speech Recognition”. *Proc. IEEE ICASSP-99*, pp. 729-732.

[6] G. Bordel, I. Torres, E. Vidal. “Back-off Smoothing in a syntactic approach to Language Modelling”. *Proc. ICSLP-94*, pp. 851-854.

[7] I. Torres, A. Varona. “An efficient representation of k-TSS language models”. *Proc. of the IV Simposio Iberoamericano de Reconocimiento de Patrones, La Habana, Cuba, march 1999, pp. 645-654.*

[8] L.J. Rodríguez, M.I. Torres, J.M. Alcaide, A. Varona, K. López de Ipiña, M. Peñagarikano, G. Bordel. “An integrated system for Spanish CSR tasks”. *To appear in Proceedings of the 6th European Conference on Speech Communication and Technology, EUROSPEECH-99, to be held on September 5-9 1999, in Budapest, Hungary.*

[9] J.E. Díaz, A.J. Rubio, A.M. Peinado, E. Segarra, N. Prieto, F. Casacuberta. “Development of Task Oriented Spanish Speech Corpora”. *Proc. EUROSPEECH-93 (included in addendum).*