Search Space Reduction in Automatic Speech Recognition by means of Neural Networks Estimations[1]

Javier Macías-Guarasa, Juan M. Montero, Rubén San-Segundo, Roberto Barra

Speech Technology Group. Department of Electronics Engineering. UPM.
ETSI Telecomunicación. Ciudad Universitaria s/n. 28040-Madrid. SPAIN
{macias, juancho, lapiz, barra}@die.upm.es

Abstract. Traditional approaches to lower computational requirements in automatic speech recognition systems include the use of progressive search strategies and beam-search techniques. In this paper we propose a novel strategy based on the use of neural networks to carry out estimation tasks related to the reduction of the search space in both isolated word (IWR) and continuous speech recognition (CSR) systems. In the IWR case, a hypothesis module generates a list of candidate words to be forwarded to the verification stage. A neural network is in charge of deciding the size of the search space to be faced by the verification stage. The main achievement has been a statistically significant relative decrease in inclusion error rate of 33.53%, while getting a relative decrease in average computational demands of up to 19.40%. In the CSR case, we propose the use of a neural network to automatically estimate, in a frame-by-frame basis, the beam-search width to be used. In this case, the neural network based strategy has proved to be comparable to the use of an empirically estimated fixed beam width.

1 Introduction

Computational demands are one of the main factors to take into account when designing systems supposed to operate in real-time, especially when talking about public information services using the telephone network. Telephone information service providers are demanding systems and algorithms that allow them to increase the number of active recognizers to run in dedicated hardware, to be able to significantly decrease production costs.

According to this scenario, most of the research work aimed at lowering computational requirements has been centered in search space pruning techniques, usually based in beam search techniques applied to the state level [1]. We can also refer to alternative proposals for search-space reduction in HMM based systems, such as the one described in [2], based in the detection of state change points, and oriented to forced alignment algorithms in speaker verification tasks.

[1] This work has been partially funded by the Spanish Ministry of Science and Technology under contracts DPI2001-3652-C02-02 (URBANO-IVANHOE), TIC2003-09192-C11-07 (MIDAS-INAUDITO), and DPI2004-07908-C02-02 (ROBINT).

In addition to the beam-search strategy, state of the art systems are usually based in some form of progressive search [3], whereby successively more detailed (and computationally expensive) knowledge sources are brought to bear on the recognition search as the hypothesis space is narrowed down. This approach is a generalization of the hypothesis-verification paradigm, with several cascaded stages. In hypothesis-verification systems, the main concern is reducing the hypothesized search space as much as possible, and this is not an easy task, especially when low detailed acoustic models are used in the preselection stage.

Our proposal is focused in two directions: offering an alternative strategy for estimating beam-search widths in a spontaneous speech recognition task and reducing the search space generated by a hypothesis module in an isolated word recognition task. The common ground for both ideas is the use of the hypothesis-verification paradigm in the speech recognition systems, and the use of neural networks (NNs) as an estimation tool. The NN will be in charge of dynamically limiting the search space (on a per-utterance or a frame-by-frame basis) using any available system parameter.

2 Experimental setup

2.1 Speech Recognition Systems

Both the Isolated Word (IWR) and the Continuous Speech Recognition (CSR) systems are based on the hypothesis-verification paradigm.

In the IWR system [4], the hypothesis module follows a bottom-up approach in which a phonetic string build up algorithm (using context independent semicontinuos HMMs) is followed by a lexical access stage. The verification module is based on the Viterbi algorithm, using context-dependent HMMs. The latter receives a list of candidate words (sorted according to their likelihoods) generated by the hypothesis module and generates the final recognition result.

In the CSR system [5], the hypothesis module uses an integrated search approach combining context dependent continuous HMMS and bigram LM. It generates a word graph to be further rescored by the verification module. A refined beam-search strategy based on the use of two beam widths is used, and further optimization is achieved by means of an on-demand Gaussian values evaluation. In the current version, graph rescoring is based on additional information stored in a trigram language model.

2.2 Databases and dictionaries

In the IWR experiments, we have used a subset of the VESTEL database [6], composed of 9,790 utterances. VESTEL is a realistic speaker-independent speech corpus collected over commercial telephone lines. Cross-validation is applied by means of a leave-10%-out strategy, in order to increase the statistical significance of the results. For each of the 10 sub-experiments, 80% of the database is devoted to

training, 10% to development, and 10% to evaluation. The task dictionary is composed of 10,000 words.

In our CSR experiments, we have used a subset of the INVOCA database that was designed to support research and development in spontaneous speech recognition systems in air traffic control tasks. INVOCA contains spontaneous conversations between air traffic controllers and airplane pilots in the Madrid-Barajas (MAD) airport [5]. Our results will be based on the evaluation of the *clearances subset*, composed of 8.9 hours of recorded conversations (5,011 utterances), using 8 hours (4,588 utterances) for training and 0.9 hours (503 utterances) for testing, with a task dictionary composed of 994 words. Cross-validation is applied by means of a *leave-33%-out* strategy (3 sub-experiments). The word graph is generated by an *n*-best search strategy.

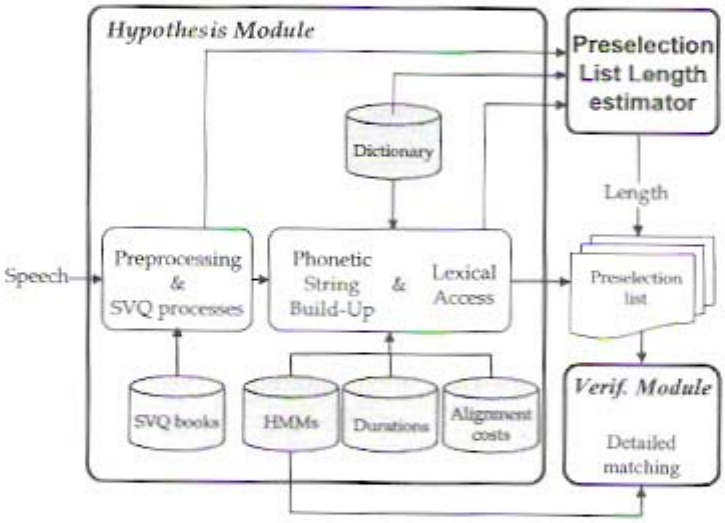

Fig. 1. Architecture for estimating PLL in the IWR task

3. Search space reduction strategy

3.1 Search space reduction in the IWR task

In the IWR system, our approach will be based on using a NN to estimate, on a per-utterance basis, the size of the list of candidate words (let us call it 'preselection list') that the hypothesis module forward to the verification stage (Figure 1). The traditional approach in these cases is using a fixed preselection list length (from now on, PLL for short), estimated according to the results obtained during system development so that a minimum error rate is achieved.

If we lower the average PLL, the computational demands of the overall system would be lower. Thus, the evaluation will be based in calculating the reduction in

average PLL (which we will refer to as *average* effort) and the relative impact in inclusion error rate.

In previous experiments, we verified that the room for improvement in computational demands in the IWR task was high, as the use of a fixed PLL while demanding a high inclusion rate lead to a very large PLL value. For example, for a requirement of 2% inclusion error rate, the average *wasted effort* is close to 94% (i.e., on average, 94% of the words in the preselection lists should not be needed if we knew the optimal PLL to be used).

In addition to that, if the neural network estimation is accurate enough, we could even get improvements in the inclusion error rate, and this actually happens in the experiments described below.

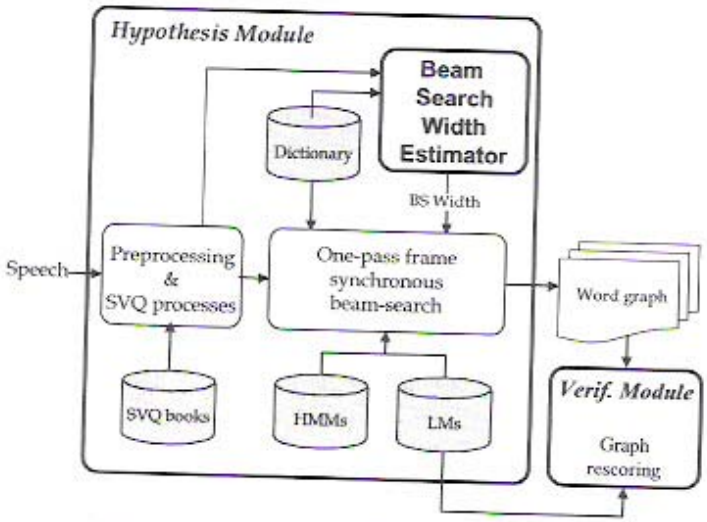

Fig. 2. Architecture for estimating BSW in the CSR task

Regarding the available input features for the NN, we designed an inventory of 56 features that can be classified in three broad classes:

- Direct parameters: Obtained from the characteristics of the acoustic utterance or the preselection process: number of frames, phonetic string length, acoustic score, etc.
- Derived parameters: Calculated from the previous ones applying different types of normalization schemes (by number of frames, phonetic string length, etc.)
- Lexical Access Statistical Parameters: Averages and standard deviations calculated over the lexical access costs distribution, for different PLLs.

3.2 Search space reduction in the CSR task

In the CSR system, our approach will be based on using a NN to estimate, on a frame-by-frame basis, the beam-search width (from now on, BSW for short) to be used in the one-pass search pruning (Figure 2). The traditional approach in these

cases is using a fixed BSW (although histogram-based beam search pruning have also been proposed in the literature, but based on histogram calculations [7]), estimated according to the results obtained during system development so that a minimum error rate is achieved.

If we lower the percentage of the full search space visited during the one-pass algorithm calculations, the computational demands of the overall system would be, again, lower. Thus, the key factor in this case is calculating the relative reduction in *visited* search space and the relative impact in WER.

In previous experiments measuring the exact BSW to be used in order to ensure minimum WER (by forced alignment and calculation, on a frame by frame basis, of the difference between the score calculated by dynamic programming and the score of the correct path), we found that the room for improvement was very low. In the third column of table 1 we show the minimum WER achievable if we knew in advance the exact minimum (optimal) BSW to use, and the percentage of the visited search space with this BSW. If we compare these values with the ones calculated by empirically estimating the empirical BSW (second column in table 1), we can see that we are very close to the optimal results, so that our proposal may not be successful in this respect.

Table 1. Comparison between using exact and empirically calculated BSW

	Empirical BSW	**Exact BSW**
WER	12.8%	11.5%
% Search space	16.7%	16.2%

In the CSR task, the design of the input feature set was similar to the IWR case, with differences due to the different characteristics of the algorithms used in the hypothesis modules. The input feature set design considered three classes:

- Direct parameters: Time index of the current frame and likelihood scores (n-best likelihood scores).
- Derived parameters: Calculated from the previous ones applying a normalization by the time index of the current frame.
- Differences of acoustic scores: The differences of scores among the ones in the n-best list, and their normalized values (dividing by the time index of the current frame)

In this initial version we did not include other possible features related to the LM (back off behavior or LM score, such as in [8]), as we were mainly interested in evaluating the potential of our proposal with a simple experimental setup.

4. Using NNs for search space reduction

4.1 NNs topology and input feature coding

In all our experiments we used a multi layer perceptron, with a single hidden layer and sigmoidal activation functions.

The input coding schemes we tested include both single and multiple inputs per input feature:

- For the single-input per parameter case, we evaluated no coding (raw data input), linearly scaling, standard z-score normalization, optional data clipping, etc.
- For the multiple-input per parameter case, we evaluated using a uniform and non-uniform distributed linear mappings.

4.2 Feature selection

In order to select the most discriminative features for our task, we used an adapted version of the greedy algorithm [9]. Initially, this procedure was designed as follows:

1. The feature set is initialized as empty.
2. In every iteration, feature ensembles are generated adding every pending feature to the existing set.
3. Experiments with variable number of epochs are performed for every feature ensemble.
4. The feature ensemble achieving the highest reduction in error rate for the optimal number of epochs is selected as the new feature set for the next iteration.
5. Continue with step 2 if the error rate improves.

Initial evaluation showed that the computational complexity of step 3 is huge, so that we simplified the process as follows:

- In steps 3 and 4 we select the 8 feature ensembles which lead to the best results using a training procedure with the optimal number of epochs found in the previous iteration.
- Before step 5 we carry out experiments to determine the optimal number of epochs for each of the 8 best ensembles and we finally select the ensemble with the highest reduction in preselection error rate.

With this approach, a set of 4 features was selected as the optimum in the IWR task (related to the standard deviation of the lexical access costs, the normalized acoustic score from the phonetic string build up module and the phonetic string length) and 4 features also in the CSR task (time frame, normalized acoustic score from the one-pass algorithm and differences between different scores in the n-best list per frame).

4.3 NN estimation in the IWR task: setup strategy

In the IWR case, our objective was directly estimating the PLL to be used. Preliminary experiments showed that using multiple output neurons was the better approach, as each output neuron could encode PLL values in a better way. Regarding output coding, using a uniformly distributed linear mapping function lead to very bad results, as only the first few neurons are activated during training, as most utterances are recognized for the first few candidates in the preselection list. The NN training setup strategy followed these two ideas:

- Every output neuron k (k=1..N) is defined to represent a different PLL range (PLLs from lowerSegmentLength(k) to upperSegmentLength(k)), leading to the task formulated as a classification problem.

- The PLL ranges that every output neuron represents are trained with a criterion that aims to get, when possible, a uniform number of training samples for all of them, in order to avoid data sparseness during training.

We also tested different alternatives regarding NN output post-processing and finally verified that the best one was based in equation (1):

$$PLL = \sum_{k=1}^{N} upperSegmentLength(k) \cdot act(k) + fxThresh \tag{1}$$

Where *act(k)* is the activation value for the *k* output neuron and *fxThresh* is a fixed threshold automatically estimated during the NN training phase. The rationale for this approach is based on the fact that, given certain premises, NN outputs can be interpreted as class posterior probabilities, so that all output neurons have something to say regarding the estimated PLL.

4.4 NN estimation in the CSR task: setup strategy

In the CSR case, our objective was directly estimating the BSW to be used.

Getting to a working NN estimation system was much more difficult in the CSR task, mainly due to the fact that the achievable improvements were mucho more limited than in the IWR case, as mentioned above. A lot of preliminary experimentation was carried out in order to validate a suitable approximation to the design of the BSW estimation. From this experimentation, we verified that using a single output neuron and estimating the BSW with equation (2) lead to the best results:

$$BSW = fxThresh + act(k) * prThresh \tag{2}$$

Where *act(k)* is the activation value for the *k* output neuron and *fxThresh* and *prThresh* are fixed thresholds empirically estimated during the system development phase.

5. Experimental results

5.1 Evaluation in the VESTEL IWR task

In the baseline experiment using fixed PLLs, we calculated the inclusion error rate achieved for every possible length of the preselection list. The inclusion error rate is obtained assuming a recognized word is within the first K candidates (K equals the PLL) proposed by the hypothesis module. Our requirement regarding inclusion rate was achieving a value under 2%. Taking into account previous experiments, we established the baseline system as the one that used exactly 1000 candidates for the fixed PLL (10% of dictionary size), which lead to an inclusion error rate of 1.72%. So, our target will be achieving at least similar performance (1.72% error rate) while, reducing the average PLL (which equals to the average hypothesized search space size, thus lowering the computational demands for the whole system), as we will use variable PLLs estimated using a neural network (NN).

In Table 1 we show a comparison of the results obtained by the baseline system (fixed PLL) and the one using a NN as a per-utterance PLL estimator (with 95% confidence intervals).

A it can be clearly seen, the relative improvement in inclusion error rate has a significant value of 33.7%, while the average effort also decreases

Table 2. Evaluation results in the IWR task

	Fixed PLL	NN based PLL
Inclusion error rate	1.72% ± 0.26%	1.14% ± 0.21%
Average effort	1000	806
relative Δ error (%)	-	-33.7%
relative Δ effort (%)	-	-19.4%

5.2 Evaluation in the INVOCA CSR task

The baseline results were calculated using a fixed BSW, calculated empirically during system development. Minimum achievable WER was 13.10%, with the search process *visiting* 14.3% of the full search space.

In Table 2 we show a comparison of the results obtained by the baseline system and the one using a NN as a per-utterance PLL estimator (with 95% confidence intervals).

As it can be seen, we get a relative (and not statistically significant) improvement in WER of 0.4%, while increasing the *visited search* around 3.5%.

Unfortunately, our NN based approach has not been able to achieve results that outperform the traditional strategy of using empirically estimated (fixed) BSWs. Nevertheless, the methodology employed is general enough to be considered for any speech recognition task, and this is the main advantage of the NN based approach.

Table 3. Evaluation results in the CSR task

	Fixed BSW	NN based BSW
WER	13.10% ± 0.82%	13.05% ± 0.81%
% of Search Space	14.3%	14.8%
relative Δ error (%)	-	-0.4%
relative Δ search (%)	-	3.5%

6. Conclusions and future work

In this paper we have proposed a novel strategy based on the use of neural networks as efficient tools in order to carry out estimation tasks related to the reduction of the search space in both isolated word (IWR) and continuous speech recognition (CSR) systems. Our proposal is to dynamically calculate parameters related to the hypothesized search space, using neural networks as the estimation module and designing the input feature set with a careful greedy-based selection approach.

In the IWR case, a hypothesis module generates a list of candidate words to be forwarded to the verification stage. A neural network is in charge of, on a per-utterance basis, deciding the size of this list, which roughly represents the search space to be faced by the verification stage. The main achievement has been a statistically significant relative decrease in error rate of 33.53%, while getting a relative decrease in average computational demands of up to 19.40%.

In the CSR case, we have proposed the use a neural network to automatically estimate, in a frame-by-frame basis, the beam-search width to be used in a one-pass synchronous beam search algorithm. In this case, the NN based strategy has proved to be comparable to the use of fixed width, but unable to outperform it.

In both tasks, the computational impact of the NN calculations is negligible when compared to the overall runtime of the preselection stage (under 0.01% of the total runtime).

We are currently working in extending the input feature set to be used in the CSR task, and integrating this methodology in a confidence estimation framework for speech understanding tasks.

References

[1] S. Ortmanns, H. Ney and A. Eiden, "Language-Model Look-Ahead for Large Vocabulary Speech Recognition", in Proc. of the International Conference on Spoken Language Processing 1996, vol. 4, pp. 2095-2098. 1996

[2] Li, Q. "A Detection Approach to Search-Space Reduction for HMM State Alignment in Speaker Verification". IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-9, n. 5, pp. 569-578. 2001.

[3] J.L. Gauvain and L. Lamel, "Large-vocabulary continuous speech recognition: advances and applications". Proceedings of the IEEE, Vol. 88, no. 8, pp. 1181-1200. August 2000.

[4] J. Macías-Guarasa, J. Ferreiros, J. Colás, A. Gallardo-Antolín and J.M. Pardo, "Improved Variable Preselection List Length Estimation Using NNs In A Large Vocabulary Telephone Speech Recognition System". in Proc. of the International Conference on Spoken Language Processing 2000, pp. 823-826. 2000.

[5] Fernández, F., Córdoba, R., Ferreiros, J., Sama, V., D'Haro, L.F. and Macías, J. "Language Identification Techniques based on Full Recognition in an Air Traffic Control Task ", in Proc. of the International Conference on Spoken Language Processing 2004. pp. 1565-1568. 2004

[6] Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. "The VESTEL Telephone Speech Database", in Proc. of the International Conference on Spoken Language Processing 1994, pp. 1811-1814. 1994.

[7] Ney, H., Haeb-Umbach, R., Tran, B.H., Oerder, M. "Improvements in Beam-Search for 10000-Word Continuous Speech Recognition", in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1992, vol. I, pp. 9-12. 1992

[8] R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, J.M. Pardo. "Confidence Measures for Spoken Language Systems" , in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2001, pp. 393-396. 2001

[9] J. Kittler, "Feature set search algorithms" in Pattern Recognition and Signal Processing, C.H. Chen, Ed., pp. 41-60. 1978.

Biography

Name: Javier Macías-Guarasa
Address: E.T.S.I.T. UPM. Ciudad Universitaria s/n.
28040-MADRID-SPAIN.
Education & Work experience: Javier received his MSEE degree and Ph.D. degrees from Technical University of Madrid in 1992 and 2001 (with highest distinction). Since 1990 he is a member of the Speech Technology Group and associate professor in the Department of Electronics Engineering in the Technical University of Madrid. He spent six months in the Speech Group of the International Computer Science Institute in Berkeley, California. His main research interests are related to Speech Technology, specially the design and development of efficient and natural man-machine dialog systems. Tel: +34-915495700 ext: 4209. E-mail: macias@die.upm.es

Name: Juan M. Montero
Address: E.T.S.I.T. UPM. Ciudad Universitaria s/n.
28040-MADRID-SPAIN.
Education & Work experience: Juan M. received his MSEE degree and Ph.D. degrees from Technical University of Madrid in 1992 and 2003 (with highest distinction). Currently, Juan M. is associate professor at the department of Electronic Engineering at ETSIT (UPM) and member of the Speech Technology Group (GTH). Tel: +34-915495762 ext: 4206. E-mail: juancho@die.upm.es

Name: Rubén San-Segundo
Address: E.T.S.I.T. UPM. Ciudad Universitaria s/n.
28040-MADRID-SPAIN.
Education & Work experience: Rubén received his MSEE degree and Ph.D. degrees from Technical University of Madrid in 1997 and 2002 (with highest distinction). Ruben did two summer stays in The Center of Spoken Language Research (CSLR-CU). From Sep. 2001 through Feb. 2003, Rubén worked at the Speech Technology Group of Telefónica I+D. Currently, Rubén is associate professor at the department of Electronic Engineering at ETSIT (UPM) and member of the Speech Technology Group (GTH). Tel: +34-915495762 ext:4228. E-mail: lapiz@die.upm.es

Name: Roberto Barra
Address: E.T.S.I.T. UPM. Ciudad Universitaria s/n.
28040-MADRID-SPAIN.
Education & Work experience: Roberto received his MSEE degree from Technical University of Madrid in 2005 (with highest distinction). Roberto is a Ph.D. student at the department of Electronics Engineering at ETSIT (UPM), with an official grant from the Spanish Government. He is also a member of the Speech Technology Group (GTH) where his main research interests are related to Speech Technology and natural human-machine spoken interfaces, with emphasis in emotional speech. Tel: +34-915495762 ext:4241. E-mail: barra@die.upm.es