ACOUSTICAL AND LEXICAL BASED CONFIDENCE MEASURES FOR A VERY LARGE VOCABYLARY TELEPHONE SPEECH HYPOTHESIS-VERIFICATION SYSTEM

J. Macías-Guarasa, J. Ferreiros, R. San-Segundo, J.M. Montero and J. M. Pardo

Grupo de Tecnología del Habla. Departamento Ingeniería Electrónica. Universidad Politécnica de Madrid.
E.T.S.I. de Telecomunicación, Ciudad Universitaria s/n, 28040, Madrid, Spain
e-mail: {macias, jfl, lapiz, juancho, pardo)@die.upm.es

ABSTRACT

In the context of large vocabulary speech recognition system, it's of major interest to classify every utterance as being correctly or incorrectly recognised.

In this paper we are presenting a preliminary study on a word-level confidence estimation system based on the output of a neural network. We use a combination of multiple features extracted from the acoustical and lexical decoders of our reference system, those available in the hypothesis stage of a hypothesis-verification very large vocabulary telephone speech recognition system. We will show the system architecture, describe the experiments leading to the selection of the set of parameters to be used by the NN and the final performance, showing promising results as compared with the use of standard log-likelihood ratio techniques for confidence scoring.

1. INTRODUCTION

Traditionally, automatic speech recognition systems rank the output hypothesis computing certain scores for each of them but they do not offer any reasonable indication of to what extent we can be confident on the proposed decoding [3].

For applications to react and prevent undesirable errors and user frustration situations, they must be able to assess the confidence that the input has been decoded correctly [5]. A lot of work is being invested in developing confidence measures for spoken language systems using acoustic [5][8] or linguistic features [1][3]. The methods used range from direct estimation using simple thresholds to LDA models for parameter selection [8] and neural networks [9].

Typical applications range from simple utterance verification, to OOV and keyword spotting.

Acoustic features alone typically show poor results as confidence estimators [5], although improvements have been achieved when using word lattices as information sources. The main reason for this poor behaviour relies in the fact that the likelihoods generated by the acoustic decoders (and even when combined with the language model) are relative to the probability of the observed acoustic themselves, so that they are not comparable across utterances, which is an essential feature of any confidence measure [7].

The confidence measure literature is traditionally centred in the description of methods to convert the likelihoods and joint probabilities offered by HMM decoders into useful confidence measure with different success stories. Different types of normalisation or the application of post-classifiers are typical strategies in this area [7], using parameters including likelihoods [5], LM probabilities [1] and information from n-best decoding lists [7].

In our case we will make no assumption in advance, testing every known-in-advance system feature.

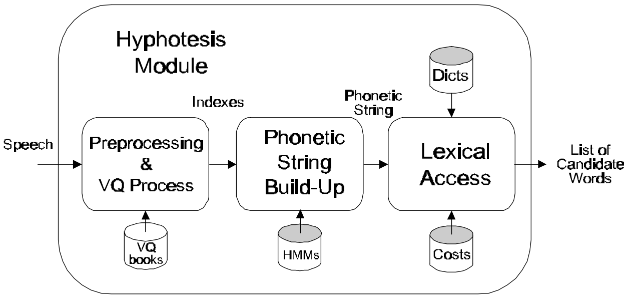

Figure 1: Hypothesis system architecture.

2. RECOGNITION SYSTEM OVERVIEW

The current implementation of the hypothesis module follows a bottom-up, two-stage strategy, as shown in Figure 1. The main preselection modules [2][10] are:

Pre-processing and VQ: The front end processing obtains an 18-dimensional vector composed of 8 MFCCs, 8 delta-MFCCs, cepstral energy and its first derivative. It performs a quantisation process for discrete HMMs (DHMMs) or soft quantization if semi-continuous HMMs (SCHMMs) are used, with up to 2 codebooks and 256 centroids each.

Phonetic String Build-Up (PSBU): the resulting indexes are passed on to the phonetic string build up module, which generates a string of alphabet units. We have selected the One-Pass algorithm with minor modifications, using 25 allophone-like units that have been automatically extracted using an entropy based HMM clustering algorithm. The initial allophone inventory was composed of 51 units, 2 of them devoted to modelling initial and final silences. We have kept the number of units so low to reduce the computational complexity required for the task.

Lexical Access (LA): The phonetic string is matched against the dictionary, using a dynamic programming algorithm and alignment costs for unit substitution, insertion and deletion errors

3. MOTIVATION

The work presented here started originally as a side-product of different studies we were doing in our laboratory in order to develop a variable preselection list length estimation system [10]. The idea was designing a system to estimate the number of words that the hypothesis module should pass to the verification stage. Given this objective, we can think that if the system works correctly, it would propose a small list if the word has been correctly recognised in the first positions of the preselection list, and viceversa. So, we decided to apply the same system in order to design a word-level confidence estimation module. Initially, we have applied these ideas to the output of the preselection system only.

4. EXPERIMENTAL SETUP

In our experiments we have used part of VESTEL, a realistic telephone speech database, captured using the Spanish public switched telephone network [6]. The VESTEL subset we have used defines three main working parts:

- PRNOK5TR: Devoted to generic system training, it's composed of 5820 files (3011 different speakers)
- PERFDV: Devoted to testing and originally designed to make "vocabulary dependent tests". It's composed of 2536 different utterances (2255 different speakers).
- PEIV1000: Devoted to testing and originally designed to make "vocabulary independent tests". It's composed of 1434 utterances (1351 different speakers)

In the tasks described in this paper, we have used dictionaries composed of 10000 words.

5. STANDARD CONFIDENCE ESTIMATION FEATURES

We ran several experiments in order to show the ability of the typical log-likelihood related features in confidence estimation in the task under study.

We extracted the values of the following features for all the databases under study:

- Acoustic log-likelihood
- Log-likelihood normalised by word frame length
- Lexical access cost for the first candidate (similar in nature to acoustic log-likelihood)
- Standard deviation of the lexical access costs, measured over the first 10 candidates in the preselection list

To evaluate the discrimination power of all of them, we calculated the statistical distributions for the two target populations: correct and wrong words (word recognised in the first position or above, respectively). As an example, we include the distribution plot for the PRNOK5TR list and the normalised log-likelihood in Figure 2 (the distributions for the other features are very similar to this one).

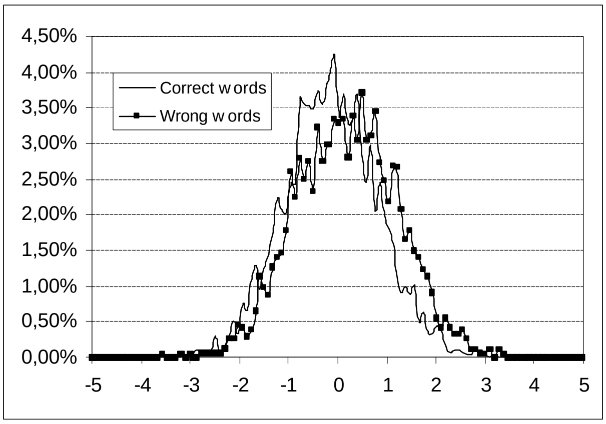

Figure 2: Distribution of the normalised log-likelihood feature, for correct and incorrect words in the PRNO5TR database

It's obvious that any discrimination approach using parameters with such statistical distribution would fail in the confidence estimation task. So it's clear we need some strategy to extract meaningful information from the raw parameters selected, and it's here where the NN is to be used.

6. NEURAL NETWORK ESTIMATOR

The neural network used is a single-output typical three-layer MLP, trained to give a high output activation value if the word is not in the first position of the preselection list, and a low output activation value if it is actually recognised in the first place. A wide range of topology alternatives and coding schemes for the input layer is to be evaluated, in order to find the best combination.

6.1. Features

Neural networks are able to achieve amazing results in really difficult tasks, but this highly depends on the quality and significance of the available input parameters, the selected coding scheme and, obviously, of the availability of enough training data.

One of the crucial questions to face when selecting the inventory of input parameters is whether the problem will be solvable with them. In our case this is not clear at all, given the lack of explicit and implicit a priori information on the estimation task, and given the discouraging feature statistical distributions shown above. So, we have created a wide spectrum of possibilities. Summarising, we have selected 33 parameters than can be classified in the following three broad classes:

- Direct parameters: Directly obtained from the acoustic utterance or the preselection process: number of frames, phonetic string length, acoustic search log-likelihood, number of symbols in the first candidate in the preselection list, lexical access cost for this candidate, etc.

- Derived parameters: Calculated from the previous ones applying different types of normalisation (dividing by number of frames, phonetic string length, etc.): acoustic log-likelihood normalised by number of frames or string length; lexical access cost normalised by the number of frames, the string length or the number of symbols; phonetic string length normalised by number of frames, etc.
- Lexical Access Statistical Parameters: Some studies in the literature used the differences in log-likelihood between the best and second best hypothesis [6] or the utterance likelihood and the one calculated using an "alternative recognition network". We decided to generalise this idea and to include certain parameters related to the statistical distribution of the lexical access costs, calculated for different preselection list lengths. So, we extracted averages and standard deviations of the costs, normalised or not, for list lengths equal to 0'1%, 1%, 10%, 25% and 50% of the dictionary size.

6.2. Topology, coding scheme and feature selection

We launched experiments for all 33 available parameters, using all available coding techniques and topology alternatives, which sums up to almost 2800 experiments with their corresponding results. The conclusions we obtained can be summarised as follows:

- The maximum rates achieved are around 70-75%, for the three databases, which is really high taking into account the simplicity of the proposed discrimination system and given the low rates obtained for the preselection system in the first candidate (46.95%, 30.14% and 42.47% for PRNOK5TR, PERFDV and PEIV1000, respectively).
- The best absolute results for the training set were obtained with the multiple input neurons per parameter and non-linear mapping when coding the inputs and using thermometers, but the differences are not statistically significant when compared with the best single input neuron methods. This means that the main factor is the information kept in the parameter itself, being less important the coding approach used. So, we decided to establish the network topology 1+5+1, using one input neuron and coding the parameters using simple normalisation in the rest of the experiments.
- The list of parameters with higher discrimination power is consistent across all databases, confirming again the fact that the network is actually able to extract the discrimination information it needs to perform well.
- The best parameter in all the experiments has been the standard deviation of the lexical access costs, measured over the list of the first 10 candidates in the preselection list (0'1% of the size of the dictionary). This is really a very important fact, as it shows that the parameter with the highest discrimination power is related to the dispersion of the lexical access costs. Additionally, it confirms experimental evidence we had in our Group related to the relationship between dispersion measures in acoustical and lexical costs and the recognition confidence (also in agreement with the results obtained by other research groups [6]).
- In contrast, the parameters related to the average of lexical access costs have proved to be very bad in the discrimination experiments, showing that the parameters related to absolute values are useless in our task.
- The parameters related to acoustical log-likelihood had shown very poor discrimination power, in all cases, hardly reaching the 55%.
- The parameters related to lexical access cost of the first candidate get reasonable good results when they are normalised, getting rates up to 64%.
- The parameters related to word length showed uneven performance, getting rates in the vicinity of 60%.

After this study on single parameter discrimination power, we developed an objective procedure to build the final feature inventory, based on progressively adding the features leading to the higher improvements in discrimination performance. After this, we came up to a list of 8 features, the ones to be used in the final system

7. EVALUATION

In this section we will give detailed figures on the performance achieved by the final system. Using 8 parameters, 5 neurons in the hidden layer and simple normalisation for the input parameters.

As we did in Section 5, we show in Figure 3 the statistical distribution obtained of the NN output for the two populations of interest:

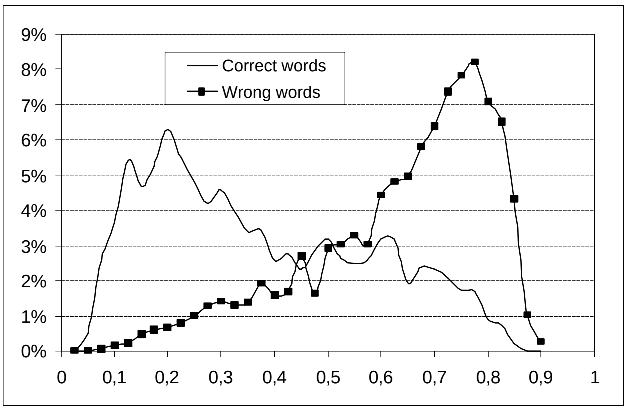

Figure 3: NN output value distribution for correct and incorrect words in the PRNO5TR database

Obviously, the discrimination task using the typical threshold based decision is much easier in this case. To give a final idea on the system performance, we include the curve of false rejection and false acceptation rates as a function of the decision threshold used in the discrimination process. In Figure 4 we show the results for the PERFDV database and in Figure 5, the ones for the PEIV1000 case.

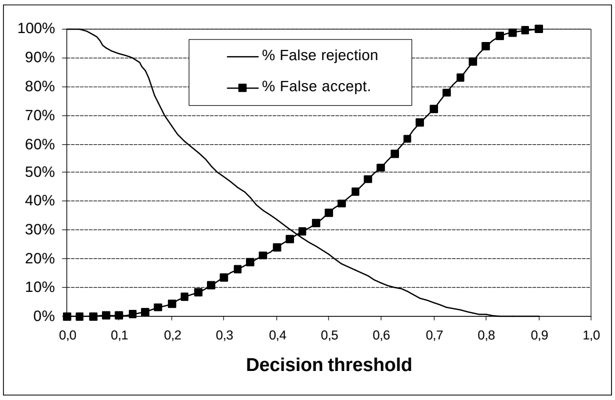

Figure 4: FR and FA rates for the PERFDV database

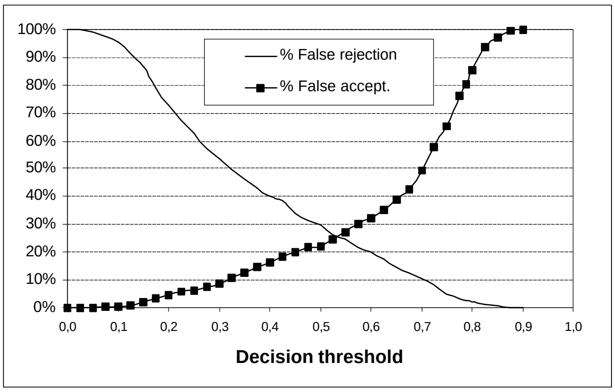

Figure 5: FR and FA rates for the PEIV1000 database

The EER is around 30% for the PERFDV case and 25% for the PEIV1000 one, reasonably high values taking into account the simplicity of the estimation system. Additionally, it's interesting to notice that the threshold for EER is very close to the 0.5 value, which further eases the discrimination process and validates our approach.

Additionally, we could analyse the effect of allowing a certain false rejection rate, finding out to what extent we can increase the correct rejection rate. In Table 1 we show the latter figure for two values of false rejection rates.

Dataset	FR=5%	FR=2.5%
PRNOK5TR	35.27 %	23.89 %
PERFDV	27.56 %	17.40 %
PEIV1000	34.70 %	19.47 %

Table 1: Correct rejection rates for given false rejection rates.

The results obtained are reasonable, especially when considered the high preselection error rates we face in the hypothesis subsystem.

8. CONCLUSIONS AND FUTURE WORK

We have introduced and evaluated a word-level confidence estimation system based on the use of neural networks and a combination of lexical and acoustical features. The original features, when used with traditional threshold-based discrimination techniques were unable to achieve good results, while the NN approach is actually able to achieve very good results given the recognition task under study.

The main work to be done is extending the confidence estimation measure to the verification module and applying this approach to continuous speech recognition systems.

9. REFERENCES

1. San-Segundo R., Pellom, B. and Ward, W. "Confidence Measures for Dialogue Management in the CU Communicator System". ICASSP 2000
2. Macias-Guarasa, J., Gallardo, A., Ferreiros, J., Pardo, J.M. and Villarrubia, L. "Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment". ICSLP 96
3. Chase, L.. "Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition". EUROSPEECH 97
4. Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. "The VESTEL Telephone Speech Database". ICSLP 94.
5. Bergen, Z. "A Senone Based Confidence Measure for Speech Recognition". EUROSPEECH 97. .
6. Dolfing, J.G.A. and Wendemuth, A. "Combination of Confidence measures in Isolated Word Recognition. ICSLP'98.
7. Williams, G. "A Study of the Use and Evaluation of Confidence measures in Automatic Speech Recognition". Technical Report C-98-02. Department of Computer Science. University of Sheffield. 1998
8. Kamppari, S.O. and Hazen, T.J. "Word and phone level acoustic confidence scoring. ICASSP2000.
9. Weintraub, M. Beaufays, F., Rivlin, A., Konig, Y. And Stolcke, A. "Neural network based measures of confidence for word recognition". ICASSP 1997
10. Macías-Guarasa, J., Ferreiros, J., Colás, J., Gallardo, A., and Pardo, J.M.. "Improved Variable Preselection List Length Estimation Using Neural Networks in a Large Vocabulary Telephone Speech Recognition System". In these proceedings.