Application of Confidence Measures for Dialogue Systems through the Use of Parallel Speech Recognizers

David Pérez-Piñar López, Carmen García Mateo

Departamento de Teoría de la Señal y Comunicaciones
University of Vigo, Spain
dperez@gts.tsc.uvigo.es, carmen@gts.tsc.uvigo.es

Abstract

To assess the correctness of a recognizer output in any instance of a dialogue is a complex task that has been studied thoroughly during the past decade. Its importance relays on the need for robust dialogue systems, capable of dealing with difficulties inherent to human-machine communications: user errors and corrections, speech recognizer errors, error recovery techniques, etc.

In this paper, we present a novel approach to the problem of deciding what the user has said. We use confidence measures derived from low level knowledge sources (acoustic and linguistic information) and generated in parallel from several topic-adapted speech recognizers. Each recognizer is aimed to the recognition of a particular topic, and confidence measures are compared through the use of a classifier that lead to a most probable solution.

This approach shows to be specially suited for difficult topics, such as proper names or confirmations, which are highly meaningful for error correction tasks. These topics present high error rates when using an application-wide speech recognizer, but recognition correction is greatly enhanced through the use of parallel recognizers. Moreover, the use of topic-adapted recognizers seems to help also in the identification of the user intention and in the detection of out-of-application utterances.

1. Introduction

Spoken Dialogue Systems (SDS) have been object of thorough research in the past decade as a result of their deployment in real-world applications: weather information services, travel information and reservation, telephone mail access and management, emergency services, etc. Most of these systems use the voice as the only communication channel. This fact limits their capabilities, and as a result they are likely to fail in some situations. Error correction techniques are therefore needed to guarantee a minimum of user satisfaction.

However, these techniques are difficult to implement. The dialogue manager needs a huge amount of information and complex algorithms to be able to solve these situations successfully without causing frustration on users.

A better practical approach is to simplify the dialogue manager module that deals with errors and to take advantage of low and high level knowledge sources, which can give many clues to detect the correctness of user utterances. Acoustic and linguistic evidences extracted from the speech recognizer module can give hints on user intention and recognizer output correctness. High level information, such as semantics, pragmatics or dialogue history, should be used to complement and reinforce hypothesis obtained previously.

These hints are traditionally generated as confidence measures, which have been object of a considerable research activity [1]. They can be derived using side-information from the recognizer, such as likelihoods [2], different decoding outputs [3], N-best lists [4], etc. In some practical cases (for instance, when using proprietary recognizers), it may be difficult to obtain these features, and therefore other methods for generating confidence measures are used, such as free phonetic recognizers, dynamic alignment, etc.

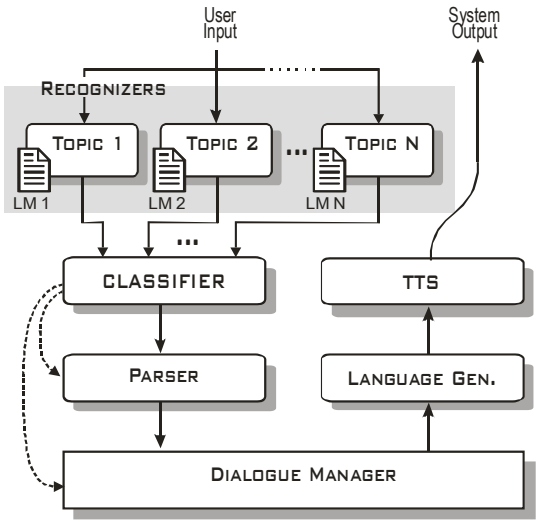

Figure 1: Dialogue System diagram with parallel input recognizers adapted to application topics

Figure 1 shows a general dialogue system diagram, where confidence measures are extracted from linguistic and acoustic features generated by parallel topic-adapted recognizers. Each recognizer specializes in the recognition of one topic through the use of mixed language models. For a specific application, each utterance is processed using as many recognizers as topics are included in the application.

The features used for the generation of confidence measures are kept simple in order to explore the possible use of this approach on systems based on third party recognizers, used by developers as black boxes. A classifier is applied at the output to determine the most probable topic for each uttered sentence, and to select the recognized output that will be used by the parser and dialogue manager.

The rest of the paper is organized as follows: next the experimental framework is presented. Section 3 describes the details of the speech recognition and classification systems. Results are presented in section 4, leading to the conclusions.

2. Experimental Framework

2.1. Baseline Recognition System

We employ a large vocabulary continuous speech recognizer based on Continuous Hidden Markov Models (CHMM). The recognition engine is a two-pass recognizer: a Viterbi algorithm, which works in a synchronous way with a beam search, and an A* algorithm [5].

Acoustic models are generated from Galician and Spanish SpeechDAT databases. These speech corpora were recorded through the public fixed telephone network, sampled at 8 KHz and codified by the A-law using 8 bits per sample. As training data we have used 15 hours in Galician and 25 hours in Spanish. From these data we generate 627 acoustic units, which are demiphones consisting of 2-state HMMs. Each HMM-state is modeled by a mixture of 4 to 8 Gaussian distributions with a 39-dimensional feature space: 12 mel-frequency cepstrum coefficients (MFCC), normalized log-energy and their first- and second-order time derivatives.

There is no need to obtain adapted acoustic models, as we are aiming for a general speaker-independent dialogue system. Trigram language models are trained using the SRILM toolkit [6] with Katz smoothing.

2.2. Test Database: Speech and Test Data

We have used a subset of the Spanish version of SpeechDAT database [7] to train, evaluate and test the recognition output classifier. The database is made up of 5,000 telephone calls. Each session comprises 14 different topics, ranging from natural speech to dates, numbers and confirmations.

This is not a dialogue database: each speaker answers questions in a poll fashion, where an automated voice server posts the question and records the audio from the user. However, our first goal is to carry out tests on the possibility of using parallel recognizers, and SpeechDAT gives us enough topics to deal with. Besides that, working with low level knowledge sources does not impose constraints related to dialogue flow, and isolated utterances can be used.

To explore all the possible scenarios, a set of topics were selected from the database:

- **Dates**. This topic includes all utterances expressing dates, in a fixed way (*November the third, nineteen ninety six*) or using natural speech (*next Monday*).
- **Names**. Proper names, usually with first and last name.
- **Numbers**. Includes telephone numbers, credit card numbers and others, expressed in natural speech (*sixty three thousand five hundred and twenty*) or using isolated numbers (*six three five two zero*).
- **Confirmations**. Utterances confirming or rejecting some information.

These topics have distinctive linguistic and acoustic characteristics, but some of them share common expressions and vocabulary (for example, dates and numbers). They are widely used in real-life dialogue systems, and specifically confirmations play a crucial role in dialogue management.

Finally, a dataset is built up by joining the four previous sets in order to model the whole speech application. Therefore, our experimental dialogue application is comprised of four possible topics, and the dialogue manager is supposed to manage them properly. Any other topic is not supported by the system.

Topic	#Training files	#Validation files	#Test files
Dates	2248	375	375
Names	1494	249	249
Numbers	3745	625	625
Confirmations	1478	247	247

Table 1: Number of (speech and transcriptions) files for each topic and partition

Orthographic transcriptions from utterances for each topic are divided into three partitions: training, validation and test. The SpeechDAT subset used contains a total of 991 speakers, 479 males and 512 females, which are assigned to different partitions. Training partition is used for language model training, contains 75% of transcriptions available for the topic at hand and excludes utterances with mispronounced or incomplete words and with intermittent noise. The validation partition is used for training the classifier, and test utterances will be used to evaluate the recognizer performance. Table 1 shows the utterance distribution for each partition and topic.

3. In-Parallel Speech Recognition

Traditional dialogue systems are based on a single recognizer with an universal language model, usually adapted to the system application. In our proposal, we divide the recognition module in several topic-adapted recognizers, whose results are combined by a post-classifier that selects the right one based on confidence measures. Figure 2 shows the recognition module with the acoustic and linguistic resources used.

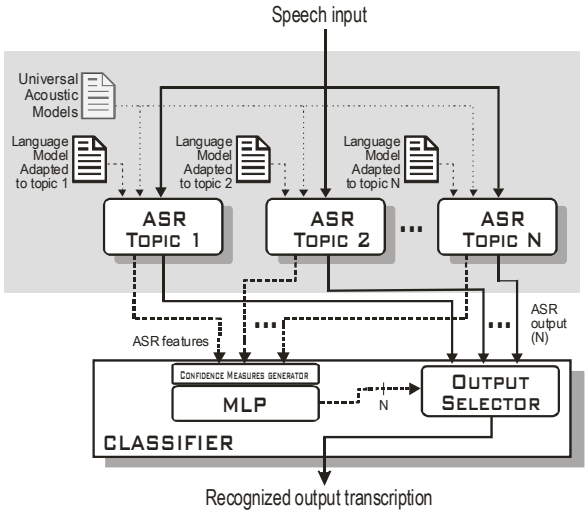

Figure 2: Recognition module diagram

3.1. Topic-adapted language models

Our topic-adapted recognizers make use of topic-adapted language models generated from SpeechDAT orthographic transcriptions. Topic adaptation is achieved by mixing n-gram

models [5]. This mixture is generated in several steps. First, separate trigram language models are trained for each topic and for the whole application, using Good-Touring discounting and backoffs. The original topic vocabulary has 115 words for dates, 581 for names, 99 for numbers and 66 for confirmations.

The application vocabulary contains 761 words. An additional universal language model is obtained from newspaper corpora, with a 20k vocabulary size. After mixing, vocabulary size is approximately 20k words for every LM.

For the sake of simplicity, in our experiments mixture weights are fixed and equal to 15% of topic LM and 85% of universal LM. A higher relative weight of the topic language model (for instance, 30%) will give better results for topic utterances, but will also give higher out-of-vocabulary hits. A lower weight (for instance, 5%) will reduce out-of-vocabulary word rate, but the resulting mixed LM will be too general to discriminate between topic utterances and out-of-topic expressions. A more precise solution would select the weights that minimize perplexity for each mixed language model on an evaluation set like done in [5].

3.2. Confidence measures

Each recognizer generates some useful features as decoder side-information: acoustic likelihoods, language model probabilities, N-best results for each utterance, time-alignment information, etc. This information is processed to generate confidence measures at the word and phrase levels.

While deciding what features to use and what feature combination to apply, our goal is two-folded. Evidently, we need meaningful features for the task at hand. But, as stated previously, we will try to use commonly available features to test the performance of our approach for systems using third party recognizers, which usually cannot be extended to generate additional features.

Therefore, in our preliminary experiments we will only use acoustic likelihoods, linguistic probabilities and decoded transcription, generated directly by our recognizer. Three confidence measures are derived from these features:

- Normalized Sentence Acoustic Score (NSAS). The acoustic likelihood of each word in the recognizer output is summed up and normalized with respect to the number of recognized words (NRW).
- Normalized Sentence Linguistic Score (NSLS). The language model probability associated with each word is summed up and normalized in the same way as the previous score.
- Number of Recognized Words (NRW). It is just the number of words in the recognizer output.

The relation of these features with recognition correctness can be shown by correlating their values with recognition assessment results. As an example, figure 3 shows NSAS distribution using the dates-adapted recognizer for two different topics: dates and names. Results are promising, as distributions are clearly different.

This preliminary assessment was carried out for every possible combination of recognizer and topic, and applied to NSLS also. Results demonstrate that different topics convey different distributions. They are, therefore, separable classes, and a statistic classifier can be trained to learn each topic and help the dialogue manager to distinguish them.

3.3. Recognizer output classifier

Confidence measures are fed into a neural network that selects the most probable recognizer output. Simultaneously, classification gives the dialogue manager information about the user intention in each dialogue step.

Many different classification techniques could be used for this task [3]. After several tests, a single Multi-Layer Perceptron (MLP) was selected as a compromise between training performance and classification accuracy. The network receives confidence data from each recognizer, which means 12 inputs. The hidden layer contains 12 neurons of tanh non-linearity type. Finally, four binary output cells encode the topic detected.

MLP training is done with validation data extracted from SpeechDAT. Two important details should be noticed:

- While language model probability shows a very stable behavior, the recognizer generates some acoustic likelihoods that are clearly incorrect. These outlayers should be discarded to avoid distortion while training.
- In highly non-linear systems, like the MLP, the order in which training data is seen is important, especially when the number of training patterns is not great. Data must be randomized before each training session to avoid low frequency regularities to be fed into the network.

In order to cope with the variability that shows up when training the neural network, a simple genetic algorithm is implemented. This algorithm optimizes the weights and the number of hidden processing elements, executing consecutive training steps and selecting the optimum solution in the sense of minimum output error. This optimization algorithm has shown to be effective [9], reducing the overall error in our experimental framework in 4.2%.

4. Experimental results

Before defining and training the neural network, a preliminary recognition assessment was carried out using the test partitions defined previously. Table 2 shows the results of recognizing

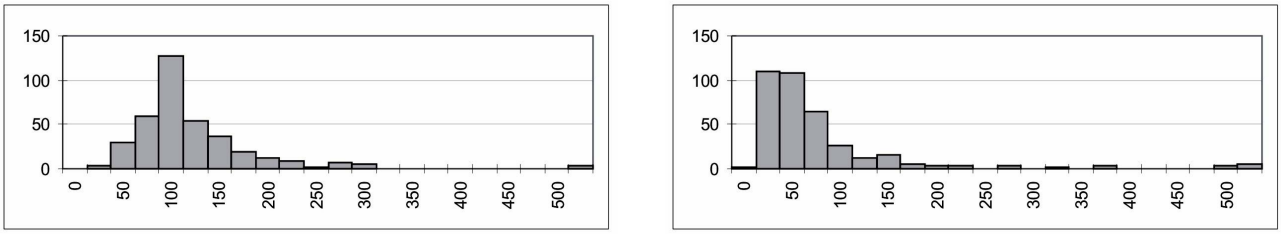

Figure 3: NSAS distribution for dates-adapted recognizer applied to dates (left) and names (right)

each topic (columns) using each topic-adapted recognizer (rows). Table 2-a shows recognition rate percentage, and Table 2-b shows % recognition improvement using the universal language model results as a baseline.

LM	App.	Conf.	Dates	Names	Numbers
universal	75,2	40,0	75,1	41,6	79,3
application	**88,5**	57,5	88,2	82,1	90,7
confirmation	68,8	**79,6**	60,7	41,9	73,9
dates	78,3	30,4	**88,1**	40,7	78,4
names	63,9	32,9	57,7	**82,9**	67,4
numbers	81,0	26,4	71,6	41,6	**90,9**

Table 2-a: % Recognition rate

LM	App.	Conf.	Dates	Names	Numbers
application	**13,3**	17,5	13,1	40,4	11,4
confirmation	-6,4	**39,6**	-14,3	0,2	-5,4
dates	3,0	-9,6	**12,9**	-0,8	-0,9
names	-11,2	-7,1	-17,3	**41,3**	-11,8
numbers	5,7	-13,5	-3,4	0,0	**11,6**

Table 2-b: % Recognition improvement

Results show a great improvement in recognition correction for the individual topics and also for the application. Results are especially good for difficult topics, such as names and confirmations. The universal model generated from news corpora does not model these cases correctly, but topic-adapted recognition does. Therefore, this approach is well suited for dialogue systems, where these topics are common.

Another important result is that utterances corresponding to a particular topic are best recognized with the recognizer adapted to that particular topic, and other utterances are harder to recognize. This means that the classifier will be able to identify the right topic and to select the right transcription.

In our second experiment, we use the MLP described before to select the recognized output for each case. The test partitions are joined and randomized, then fed into the recognizer. Features from each topic-adapted recognizer are combined and fed into the MLP, and the network output indicates the most suitable topic. Table 3 shows the overall results as a confusion matrix, with test input topics as rows and classifier output as columns.

Input topic	Conf.	Dates	Names	Numbers
confirmation	**74,1**	10,7	1,5	13,7
dates	4,9	**87,7**	5,7	1,7
names	0,7	6,1	**89,8**	3,4
numbers	8,5	6,7	2,5	**82,3**

Table 3: Classifier confusion matrix

Recognition performance is greatly enhanced for confirmations and names when compared to our universal baseline. Results are also a bit better for dates and numbers, but improvement is lower in these cases. From a dialogue management point of view, results are also encouraging, as the recognition module will select the correct topic with an overall error of 16,45%. However, more experiments should be carried out, as these results have shown to be highly dependent on the neural network structure and training.

5. Conclusions

A new approach to the design of recognition modules for dialogue systems has been presented. Its performance is better than traditional universal recognition, especially for those cases that are not well modeled, such as proper names and confirmations. These topics present a crucial importance for dialogue systems, as they are commonly used and (in the case of confirmations) play an important role in error detection and error correction techniques.

Performance should be improved further if language model mixture weights are optimized by minimizing LM perplexity using the EM algorithm. And using high level confidence measures will help the classifier to solve some of the incorrect cases shown before. These measures can be easily generated from syntactic and semantic features extracted by the parser, and are suggested as a main research line to complement these results.

6. Acknowledgements

This project has been partially supported by Spanish MEC under the project TIC2002-02208, and Xunta de Galicia under the projects PGIDT03PXIC32201PN.

7. References

[1] Cox, S. and Dasmahapatra, S., *High-level Approaches to Confidence Estimation in Speech Recognition*, IEEE Trans. on Speech and Audio, 7(10):460-471. November 2002.

[2] Schaaf, T. and Kemp, T., *Confidence Measures for Spontaneous Speech Recognition*. In Proc. IEEE Conf. on Acoustics, Speech and Signal-processing, April 1997.

[3] San-Segundo, R., Pellom, B. et al.. *Confidence measures for spoken dialogue systems*, in Proc IEEE ICASSP, Pp 393-396, ISBN 0-7803-7041-4, Salt Lake City, 2001.

[4] García-Mateo, C., Reichl, W., Ortmanns, S., 1999. *On Combining Confidence Measures in HMM-based Speech Recognizers.* Intern. Workshop on Automatic Speech Recognition and Understanding, ASRU99, Keystone, CO, USA, December 12–15, 1999.

[5] Dieguez-Tirado, J., García-Mateo, C. et al., *Adaptation strategies for the acoustic and language models in bilingual speech transcription*. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 833-836. March 2005

[6] A. Stolcke, "SRILM – an extensible language modelling toolkit," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, Denver, CO, September 2002, pp. 901–904.

[7] Moreno, A., *SpeechDAT Spanish Database for Fixed Telephone Networks*. Corpus Design Technical Report, SpeechDAT Project LE2-4001. 1997.

[8] Chase, L. L., *Error-responsive feedback mechanisms for speech recognizers*, PhD thesis, School of Computer Science, Carnegie Mellon University, 1997.

[9] Khare, V. and Yao, X., *Artificial Speciation of Neural Network Ensembles*, Proc. of the 2002 UK Workshop on Computational Intelligence (UKCI'02), pp. 96-103, Birmingham, UK, September 2002.