Fuzzy reasoning in confidence evaluation of speech recognition

Gustavo Hernández-Ábrego and José B. Mariño*

TALP Research Center
Dept Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya
Jordi Girona 1-3, Campus Nord D-5, Barcelona 08034, Spain
(abrego/canton)@gps.tsc.upc.es

Abstract

Confidence measures represent a systematic way to express reliability of speech recognition results. A common approach to confidence measuring is to take profit of the information that several recognition-related features offer and to combine them, through a given compilation mechanism, into a more effective way to distinguish between correct and incorrect recognition results. We propose to use a fuzzy reasoning scheme to perform the information compilation step. Our approach opposes the previously proposed ones because ours treats the uncertainty of recognition hypotheses in terms of "possibility" contrasted to the "probability" notion of similar works. Experimental results over isolated words, continuous speech and keyword spotting recognition tasks show higher performance of our system compared against standard compilation methods. Here we demonstrate that, due to their approach to uncertainty; to their capabilities to handle expert knowledge and to their versatility, Fuzzy Inference Systems represent a natural way to add up recognition information into confidence measures.

1. Introduction

In spite of the multiple efforts done to date on automatic speech recognition technology, its results are not perfect. Every time a recognized word sequence is considered, there is some degree of uncertainty about its correctness. Confidence measures (CM's) represent a feasible way to express which of the recognized sequences are likely to be correct and which can be disregarded as incorrect. A rather simple technique, that has shown remarkable results, to generate confidence measures is known as "Likelihood score ratio" (LSR) [10]. It is done by normalizing the likelihood score resulting from the Viterbi decoding process by the likelihood score produced by an alternative recognition network. In our work, we add other information related to the speech recognition process to the LSR by means of a fuzzy inference system in order to build a more reliable measure of confidence. This paper is organized as follows: first, in section 2, features extracted from the recognition process are considered. In section 3, the importance of gradual terms in confidence measuring is remarked and fuzzy logic is presented as a suitable framework to deal with degrees of confidence. Some configurations for the feature-compilation step are described on section 4. The experimental frameworks in which our system is tested are described on section 5 and the results obtained are discussed in section 6. Conclusions and future lines of research are enumerated on section 7.

*This research was supported by CONACyT and by CICYT under contract TIC98-0423-C06-01

2. Features to express degrees of confidence

Confidence measures can be generated by combining information about the recognition system in a feature-compilation fashion [2]. This approach has proven to rise the discriminative power of CM's when the features are extracted from the comparison of alternative recognition hypotheses or from multiple hypotheses recognition schemes [11]. Some features, by themselves, can be directly treated as CM's, however, not every feature presents high discrimination capabilities and some of them may represent a high effort to be calculated without high performance as compensation. Through careful study of the nature of the recognition process, we have formulated three features to be the basis of our experimentation:

2.1. Likelihood score ratio (LSR)

Our first feature is the likelihood score ratio (LSR) itself. For its calculation, the likelihood score of the recognition hypothesis is normalized by the score of an alternative recognition network:

$$LSR = \log L(\vec{X}|\Lambda_p) - \log L(\vec{X}|\Lambda_a). \qquad (1)$$

$\vec{X}$ is the vector of acoustic features related to the actual input utterance and Λ_p and Λ_a are the sets of hidden Markov models of the principal and alternative recognition networks respectively. Due to its unconstrained (and inaccurate) nature, the purpose of the alternative network is to model the unrestricted signal probability, $P(\vec{X})$. This procedure tends to approximate Bayes law in posterior probability calculation. Because its simplicity and its high performance [10], we consider this feature as our baseline.

2.2. Sequence alignment score (SAS)

Our second feature is what we call "sequence alignment score" (SAS). In the calculation of likelihood score ratio, the scores of both recognizers have been considered, but the decoded strings (the main product of recognition) have been disregarded. We have reported in a previous paper [3] that a proper comparison between the principal recognition hypothesis and the alternative one can result in an efficient feature for confidence measuring. The reasoning behind this feature is to consider the alternative sequence as a "second opinion". Although this sequence cannot be considered as a recognition hypothesis, it can provide information about the nature of the recognition process. This approach is different from considering multiple hypotheses of recognition since both "opinions" are taken from two completely different recognizers. To calculate SAS, the principal recognition hypothesis is transcribed into the phonetic units used by the alternative recognizer. This "principal" string is time aligned against the "alternative" recognition string. The cost of the alignment for every pair of units (principal and alternative) is taken from a confusion matrix previously calculated that covers the typical hits, confusions, deletions and insertions present when recognition is made with the alternative units set. The overall alignment score of both strings is what we call SAS. A graphical representation of the calculation of SAS is presented on figure 1 and expressed in formulae by:

$$SAS = \max_{j(i)} \prod c(p_i|q_{j(i)}) \qquad p_i \in P,\ q_{j(i)} \in Q \quad (2)$$

where P is the principal string and Q the alternative one. $c(p_i|q_{j(i)})$ is the confusion probability of the pair of units $p_i, q_{j(i)}$. The index j is a function of i according to the restrictions shown on the right part of figure 1.

This feature tends to model the distance between two different recognition opinions. The alternative hypothesis will only be close to the principal one when both of them present similar results. In order to avoid high scores from common confusions, phonetic units have been gathered in groups under confusability criteria and penalties to the alignment score are applied accordingly. This feature, by itself, has a medium discriminative power that does not surpass the performance of LSR (see figures 4 through 6).

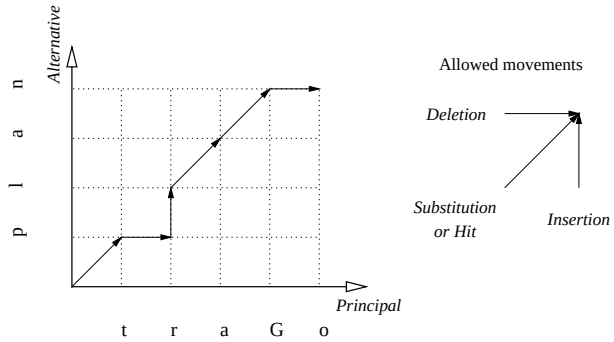

Figure 1. Calculation of SAS

2.3. Relative speaking rate (RSR)

When dealing with continuous speech, recognition errors are not just of confusion nature. Instead, insertions and deletions are a common presence. Our third feature is conceived to handle this sort of errors. We call it "Relative speaking rate" (RSR) and it is calculated according to:

$$RSR = \frac{N_T - N_{i,f}}{t_T - t_{i,f}} \quad (3)$$

where N_T is the total number of speech units (words, phones, etc) detected in the whole recognition hypothesis and $N_{i,f}$ is the number of units detected in the time interval (t_f, t_i) considered. t_T is the duration of the whole utterance and $t_{i,f}$ is the length of the time interval. This is a relative measure because, to calculate its value for a single word, it takes into account information from the whole phrase. This feature is intended to detect the lack or excess of phonetic units in a recognition hypothesis. It should be noticed that this feature is only useful on continuous speech and its use on isolated words or keywords spotting is not recommended because it may lead to wrong values of confidence.

3. Fuzzy Inference Systems in confidence measuring

Probability, understood from a frequency point of view, deals with uncertainty in terms of occurrences of known facts. In the case of speech recognition results, the known facts are whether the recognition hypothesis is correct or not. Probability is useful when dealing with serial events that require an enumeration notion of uncertainty but is not very useful when the uncertainty is about the degree of accomplishment of a known situation [5]. This is the case of confidence measuring where the task is to know *for every single recognition hypothesis*, its degree of possible correctness. The notion of "possibility" opposed to "probability" is a relevant contrast that fuzzy logic presents in front of probability theory.

The spirit of confidence measures is to express the uncertainty of speech recognition results in gradual terms and not in frequency. Under such consideration, fuzzy logic represents natural foundations for confidence measures. But fuzzy logic is not just a theoretic tool to represent uncertainty, a good share of its success is due to the several practical implementations it has. Fuzzy logic systems or Fuzzy Inference Systems (FIS) are schemes that (among other capabilities) allow to map a number of fuzzy variable inputs into a number of fuzzy outputs [7]. The mapping is done by a set of fuzzy rules that relates inputs with outputs in an "if ... then" fashion. Inputs and outputs can be represented by means of fuzzy variables able to contain language terms and fuzzy hedges. By analyzing the histograms of each of the features that we have proposed, some characteristics of them are observed and some fuzzy thresholds to separate their values when there is a correct or incorrect result are proposed. This analysis allows us to define some rules of behavior according to the correctness status of the hypotheses. The collected expert knowledge can be condensed in a fuzzy inference system. In this application, the fuzzy system can be understood as a non-linear classifier (just as a neural network) that transforms several inputs into a unique output that compiles all the information given.

4. Feature compilation schemes

It is customary that the compilation step of the information included in the recognition features is performed by means of a uniting tool based on the development of conditional probabilities. In such a way, Bayesian classifiers [1], linear discriminative analysis [11], decision trees [9] and neural networks [12] have been used as reasoning schemes to compile the involved features. We have built some classifiers based on some of the schemes mentioned to compare their performance against fuzzy systems.

4.1. Bayesian classifier (BC)

This is a rather simple classifier that maps recognition features into a confidence measure by means of a linear combination. The coefficients for such a combination are calculated from the covariance matrix of the features derived from some training data. Throughout this paper, this procedure may also be called BC.

4.2. Neural Networks (MLP)

Neural networks have been broadly used to combine recognition features into CM's. High performing results have been reported [12] and their advantages over other combination systems have been largely discussed [11]. Network topology is always a delicate issue. Remarkable results have been achieved with multi-layer perceptrons (MLP's) when trained under a back propagation framework. Simpler configurations have been preferred instead of complicate ones since performance is quite similar [11]. For our experimentation, we have chosen a feed-forward MLP with 1 hidden layer containing 4 to 6 elements, each with a hyperbolic tangent sigmoid transfer function. The input layer deals with the values of the features and the output layer deals with the CM value. The parameters of the net are adjusted in a back propagation learning phase taking examples from a training database.

4.3. Fuzzy Inference System (FIS)

Fuzzy inference system, as a classification engine, can be equipped with expert knowledge capable to separate class elements. For what our system is concerned, a rather simple configuration is considered. A Sugeno-type FIS is chosen due to its good behavior as classifier and its simplicity [4]. The number of input variables depends on the number of features used. The output variable is the value of the CM. The fuzzy rules are designed with a "reinforcement" spirit. Likelihood score ratio is treated as the main discrimination variable and the rest of the features are employed to reinforce its values. The rules of the fuzzy inference system allows to activate and deactivate the influence of the reinforcing features conveniently. Figure 2 shows an schematic representation of the FIS used to combine likelihood score ratio and sequence alignment score. The set of rules for this system is presented on figure 3

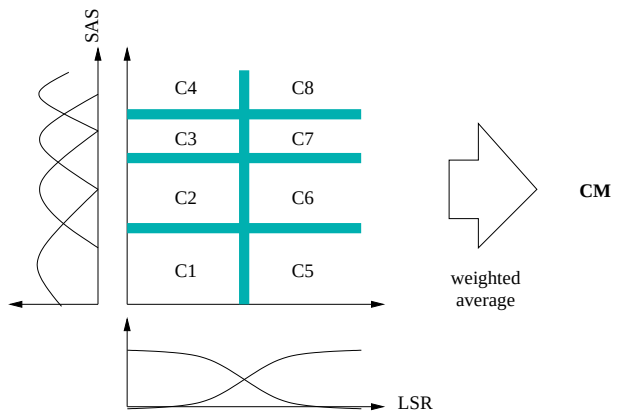

Figure 2. FIS to combine LSR and SAS

In figure 3, the consequent parts of the rules are constants, but we will prove that, even with this simple configuration, FIS performs effectively as uniting tool for CM's.

5. Experimental framework

Experimental work has been carried out with Spanish Speechdat [8] as developing and testing database. This is a database collected through the fixed telephone network,

1. if LSR = low and SAS = low, then CM	*= 0*
2. if LSR = low and SAS = midlow, then CM	*=0.05*
3. if LSR = low and SAS = midhigh, then CM	*=0.15*
4. if LSR = low and SAS = high, then CM	*=0.25*
5. if LSR = high and SAS = low, then CM	*=0.65*
6. if LSR = high and SAS = midlow, then CM	*=0.75*
7. if LSR = high and SAS = midhigh, then CM	*= 0.9*
8. if LSR = high and SAS = high, then CM	*= 1*

Figure 3. Set of fuzzy rules for a FIS of two features

sampled at 8 kHz and recorded under several acoustic environments. Speech was parameterized with mel-cepstrum coefficients. First and second order differential parameters plus the differential energy were employed. The recognition system models the phonetic units by Gaussian semi-continuous hidden Markov models (HMM's) with quantization to the 6 (2 for the energy) closest codewords. The codebook size was 256 (32 for the differential energy). Near to 1000 speakers have been selected for each of the training and testing sets. There is no speaker overlapping between sets. To cover some of the possible frameworks where CM are relevant, we have split our experimentation on the following recognition tasks:

1. Isolated words: each speaker pronounces a name of Spanish cities taken from the "City" part of Speechdat.

2. Continuous speech: speakers utter prompted time phrases. The average number of words per phrase is around 9.4. Phrases are taken from the "Time" part of Speechdat.

3. Keyword spotting: phrases containing embedded keywords are uttered by the speakers. Each sentence may contain 1 to 4 keywords. These sentences come from the "KeySentence" part of the database.

The overall task is to validate the recognition results that the recognizer produces when dealing with each of the experimentation tasks. The principal recognizer is tailored to be application independent. For acoustic modeling, it is based on high-performance sublexical phonetic units (Demiphones [6]) combined into vocabulary instances by means of an specific language model (LM). For isolated words a null grammar (all vocabulary words have equal probability) is used. Continuous speech recognition is conducted by a finite-state grammar that covers every possible time phrase. Keyword spotting has an stochastic trigram as LM. For this task, keywords are represented by Demiphones and out of vocabulary instances (OOV) by a network of phonemes. Since our purpose is to detect keywords at maximum, the presence of OOV is restricted by penalizing transitions in the phoneme network. On the other hand, the alternative recognizer is equipped with loose language restrictions and Phonemes, trained under a discriminative criterion, as phonetic units.

Detection rate, measured as the percentage of correctly recognized words, as well as the number of false alarms and the configuration of the test sets for each of the recognition tasks is shown on table 1

	isolated	continuous	keywords
voc. size	500	59	30
speakers	989	995	993
words	989	9405	1485
false alarms	172	377	1132
detection	82.61 %	95.23 %	93.80 %

Table 1. Configuration, false alarms and detection rate of the recognition tasks tested

The number of false alarms is the summation of insertions and substitutions. Without any validation of the recognition hypotheses, to retrieve the given detection rate, would imply to accept the indicated number of false alarms. Some remark about the recognition rates is worthwhile: the detection rate for isolated words is rather small due to the large number of possible words to detect. Detection in continuous speech is very high for words but not for phrases (78.19 %). In keyword spotting, it is possible to achieve a high detection rate but with a large number of false alarms as counterpart.

Generation of confidence measures passes through the calculation of features for each detected word. Next, the features of every recognition candidate are compiled, by means of one of the previously mentioned combination engines, into a CM for each word hypothesis. Performance of CM's is evaluated in terms of their capabilities to validate correct hypotheses and reject false alarms.

6. Discussion

In a task of results classification, there can be two kinds of errors: *false alarms* (i.e. wrong results regarded as correct ones) and *false rejections* (correct results wrongly regarded as incorrect). To show the relationship between these two types of errors, it is customary to generate ROC's (receiver operation characteristics). In this work, ROC's are built by varying the value of the validation threshold imposed to the resulting confidence measures and plotting the rate of false alarms against the rate of correctly recognized tokens. An ideal classifier would be able to correctly detect a large number of instances while accepting a low number of false alarms.

Figure 4 shows the ROC's of the features and combined CM's generated by different methods for the iso-

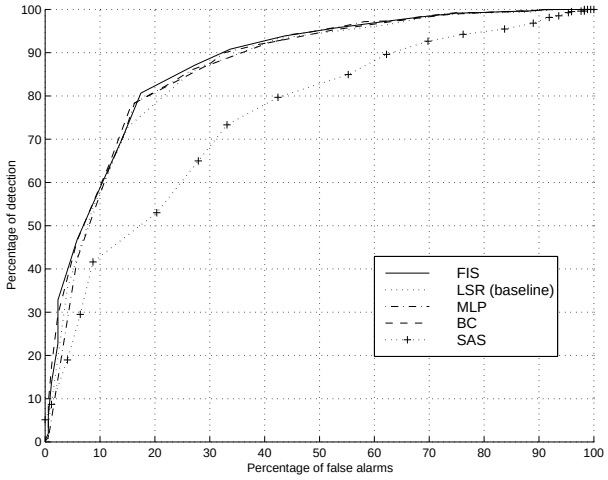

Figure 4. ROC's for the isolated words recognition task

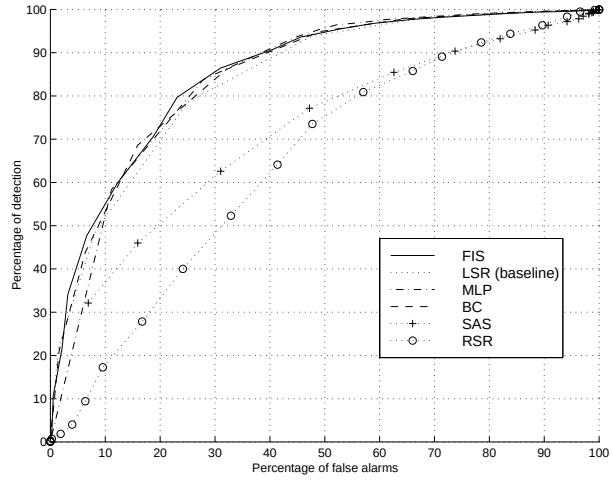

Figure 5. ROC's for the continuous speech recognition task

lated words task. As previously mentioned, LSR is considered as baseline. SAS is used as reinforcer, although it does not perform well enough in isolation. These plots show that all combination procedures present higher discriminative characteristics than the baseline. Remarkably, fuzzy system presents a good behavior along the whole plot and it is only slightly surpassed by MLP in a small region of it. Compared against the baseline, at a given false alarms rate, FIS present a noticeable increment of detection. For continuous speech, a similar behavior can be observed in figure 5. In this case, baseline has lower performance than any of the combinations. SAS has a lower performance compared to the isolated words case and here relative speaking ratio has also been considered as reinforcer, though it cannot be considered as a suitable confidence measure by itself. In this case, combinations present some irregular behavior: while BC and MLP perform well in the high detection zone, they drop in the low false alarms region. In contrast, FIS behaves satisfactorily at high detection but considerably better at the low false alarms region. For the keyword spotting case, shown in figure 6, only SAS, whose performance is quite fair, is considered as reinforcer. It is worth noticing that every combination here clearly outperforms the baseline and, among combinations, FIS shows the best performance along the whole graph.

A commonly accepted summary of ROC curves is the equal-error rate (EER). This is the point where the rate of false alarms equals the rate of false rejections. Table 2 contains the EER values for the features and the combination procedures tested. All of the combinations present important reductions of the EER compared to baseline. FIS achieves the lowest rates for every recognition task, exception made for keyword spotting where it is surpassed by the MLP at the equal error point. However, from figure 6 it can be noticed that MLP does not outperform FIS in the low false alarms region.

	Isolated	Continuous	Keywords
LSR (baseline)	20.75	23.65	21.25
SAS	30.45	34.35	31.90
RSR	-	39.00	-
BC	19.45	23.25	19.75
MLP	19.85	22.85	19.05
FIS	18.50	22.05	19.35

Table 2. EER of features and combinations

In the comparison of information compilation tools, it should be considered that the parameters of fuzzy inference systems have not been tuned up by means of the training data as the Bayesian classifier and multi-layer perceptron have, instead they have been adjusted manually based on observation. The MLP used for isolated words and the one used for keyword spotting have different parameters sets whereas the FIS remains the same on both applications. It is expected that some fine tuning procedure considering the training data (as in an ANFIS framework [4]), would rise the performance of the FIS. However, the need of specific training data to adjust parameters turns the system application dependent. So far, our compilation tools have been designed considering information extracted from each recognition task resulting in task-specific systems. This is not the case for FIS. We have efficiently applied a generic system to cover both, isolated words and keyword spotting, environments (where two features are used) without any adaptation step and we have only changed the system when a third feature is needed (in continuous speech). This versatility and transportability represents another advantage of FIS over the rest of the systems.

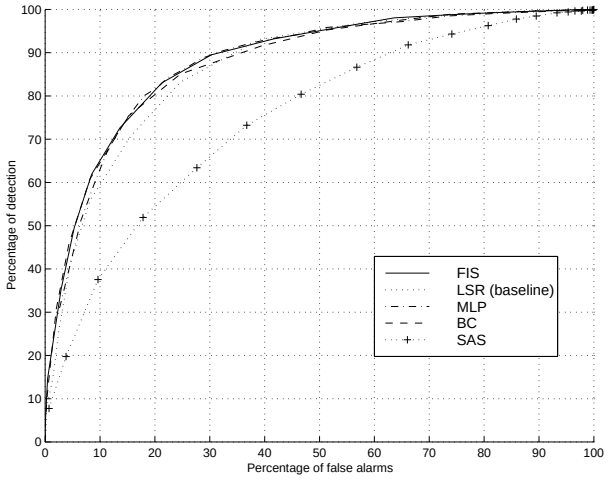

Figure 6. ROC's for the keyword spotting recognition task

Nevertheless, for every combination procedure, application dependency can be tackled by implementing sublexical (phonemes or groups of phonemes instead of words) features and confidence measures.

A final remark is deserved: the configuration of our system does not require any information derived from the recognition process but only from the recognition results by themselves. This allows to build a confidence labelers independently of the recognizer and avoids the need of exhaustive track of the whole recognition process. As a counterpart, it needs of an alternative recognition step, whose configuration is very simple, that does not represent a serious increment of the computational load.

7. Conclusions

In this paper we have shown that fuzzy logic is a natural and effective approach to measure the confidence of speech recognition results. The way it handles uncertainty, in terms of possibility, results more consistent that the way probability theory does. Fuzzy inference systems have been efficiently used to compile features related to the recognition process into a more discriminative confidence measure. They add up information into a synergetic way so the resulting combination always surpasses the original features on their own. From ROC's it is observed that the combination of features is a process worthwhile. Compared against Bayesian classifiers and multi-layer perceptrons, fuzzy inference systems show a better and more stable behavior for the recognition tasks tried, being able to maintain high detection rates while properly rejecting false alarms, even when their configuration is rather simple. Furthermore, fuzzy systems has demonstrated to be versatile and transportable between applications. The procedure we propose takes profit of information extracted exclusively from the recognition results and does not need to have any particular relation with a concrete speech recognition scheme at all.

The on-going research about this topic includes:

- to develop sublexical procedures and confidence measures in order to avoid the need of application specific data for training;
- to add information from the language model in order to improve the discrimination in continuous speech;
- to include a self-learning procedure for the fuzzy systems configuration in order to finely tune its parameters.

References

[1] S. Cox and R. C. Rose. Confidence measures for the Switchboard database. In *Proceedings of ICSLP'96*, volume I, pages 478–481, Philadelphia, October 1996.

[2] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke. Understanding and improving speech recognition performance through the use of diagnostic tools. In *Proceedings of 1995 ICASSP*, volume II, pages 221–224, Detroit, May 1995.

[3] G. Hernández-Ábrego and J. B. Mariño. A second opinion approach for speech recognition verification. In *Proceedings of the VIII SNRFAI*, volume I, pages 85–92, Bilbao, May 1999.

[4] J.-S. R. Jang. Anfis: Adaptative-network-based fuzzy inference system. *IEEE Transactions on systems, man and cybernetics*, 23(3):665–685, May/June 1993.

[5] M. Laviolette and J.W. Seaman Jr. The efficacy of fuzzy representations of uncertainty. *IEEE Transactions on Fuzzy Systems*, 2(1):4–15, February 1994.

[6] J. B. Mariño, A. Nogueiras, and A. Bonafonte. The Demiphone: an efficient subword unit for continuos speech recognition. In *Proceedings of EUROSPEECH'97*, volume III, pages 1215–1218, Rhodes, September 1997.

[7] J. M. Mendel. Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, 83(3):345–377, March 1995.

[8] A. Moreno and R. Winsky. Spanish fixed network speech corpus. Technical report, SpeechDat Project LRE-63314, 1997.

[9] C. Neti, S. Roukos, and E. Eide. Word-based confidence measures as a guide for stack search in speech recognition. In *Proceedings of 1997 ICASSP*, volume II, pages 883–886, Munich, April 1997.

[10] R. C. Rose and D. B. Paul. A Hidden Markov Model based keyword recognition system. In *Proceedings of 1990 ICASSP*, volume I, pages 129–132, Albuquerque, April 1990.

[11] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *Proceedings of 1997 ICASSP*, volume II, pages 875–878, Munich, April 1997.

[12] M. Weintraub and F. Beaufays. et al. Neural - network based measure of confidence for word recognition. In *Proceedings of 1997 ICASSP*, volume II, pages 887–890, Munich, April 1997.