A SECOND OPINION APPROACH FOR SPEECH RECOGNITION VERIFICATION

Gustavo Hernández Ábrego and José B. Mariño *

Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
Jordi Girona 1-3, Campus Nord D-5, 08034 Barcelona, SPAIN
e-mail: (abrego/canton)@gps.tsc.upc.es

Abstract

In order to improve the reliability of speech recognition results, a verifying system, that takes profit of the information given from an alternative recognition step is proposed. The alternative results are considered as a second opinion about the nature of the speech recognition process. Some features are extracted from both opinion sources and compiled, through a fuzzy inference system, into a more discriminant confidence measure able to verify correct results and disregard wrong ones. This approach is tested in a keyword spotting task taken form the Spanish SpeechDat database. Results show a considerable reduction of false rejections at a fixed false alarm rate compared to baseline systems.

Keywords: Confidence Measures, Utterance Verification, Keyword Spotting.

1 Introduction

On the purpose of compiling information useful to build confidence measures (CM), it is customary to take profit of the results obtained from the comparison of two recognition systems: the "principal" recognition network and the "alternative" one. An example of the information obtained from this comparison is the likelihood score ratio. The use of the ratio of the two recognition scores is straightforward and commonly used in the keyword spotting and utterance verification techniques [1, 2, 3]. But recognition score is not the only useful information. If well compared, the resulting word strings (that represent the main product of recognition) may give some insight into the nature of the recognition process. This sort of information has not been used, so far, for confidence measuring or utterance verification. The underlying concept about this approach is to ask for a second opinion. In every day life, when any sort of hypothesis is proposed, most of the times, we are not sure about its correctness status. However, we feel more confident about it if someone or something confirms it. Despite of the fact that this is a common procedure, it does not improve the decision taking process in all of the cases. Sometimes a second opinion just adds more hesitation to our decision and sometimes both opinions

*This research was supported by CONACyT and by CICYT under contract TIC95-0884-C04-02

may point to one direction whereas reality points to another. Nevertheless, an scheme as such can represent an useful aid to improve the performance of other knowledge sources for verifying recognition results. Features can be thereafter combined, in a synergic way, into a more powerful confidence tag helpful to classify recognition results in terms of their correctness status.

This paper is organized as follows: firstly, in section 2, we introduce the concept of "distance" between phonetic unit sequences, that can be understood as an issue of correspondence. Confusion matrices, built out from recognition evaluation, are presented as a means of expressing distance through correspondence. Based on this distance, a system that polls the opinions coming from different recognizers and that combines their results by means of a fuzzy inference system (FIS) is proposed in section 3. The overall performance of the system (in terms of discriminative power) is tested and some results reflected in section 4. Some conclusions about this method and the description of the on-going work about it appear on section 5.

2 Comparison of recognition opinions

To be able to consider a second opinion, it is mandatory to look for a measure of similarity between opinions in order to know if both opinions are coincident or not. Intuitively we can regard a pair of character sequences as "close" if several of the characters in one of them appear in the other, and as "far" if the contrary. This intuitive notion can be extended to the speech recognition framework considering two phonetic unit sequences close when they share several phonetic units or when they contain "similar" ones. We can take profit of the distance information that a confusion matrix contains in order to find the distance between phonetic sequences. From a frequency point of view, every entry of the confusion matrix can be understood as the a posteriori probability of having recognized a concrete phonetic unit given another one in the reference. Taken individually, these confusion probabilities express distance between units. Since errors and hits are not necessarily synchronous, to calculate the distance between unit sequences, it is necessary to compare all reference units against recognized ones in a methodical manner. The method employed is a dynamic programming procedure to time align a sequence against the other. As a result of this alignment we get an ideal path (ideal in terms of some predefined criterion) and an alignment score. If the confusion probabilities are used to calculate this score, a distance measure, similar to the compound confusion probability, results.

An alignment procedure assigns costs to any possible transition. This cost may vary depending on the type of transition implied and depending on which phonetic units it comprises. In an evaluation procedure, the type of transition is related to the kind of confusion occurred and the cost related to it might be different for a hit or for each of the three kinds of possible errors (substitution, insertion or deletion). By multiplying the confusion probabilities of the units involved in the best aligned path, we can calculate the total correspondence probability between the sequences involved (i.e. the score of the alignment). In formulae,

$$\begin{aligned} D_{i_{max}}(U) &= P(S|R) \\ &= P(s_j(1)|r_i(1))P(s_j(2)|r_i(2))\ldots P(s_j(k)|r_i(l)). \end{aligned} \quad (1)$$

$D_{i_{max}}(U)$ means the overall cost of the maximum path i across all sequence slots $U = u(1), u(2), \ldots, u(m)$. $S = s_j(1), s_j(2), \ldots, s_j(k)$ is the recognized sequence and $R = r_i(1), r_i(2), \ldots, r_i(l)$ the reference one. m does not necessarily have to correspond to k or l because of the insertion and deletion cases.

The alignment score should be large for sequences that contain distant (in terms of confusion probability) units and small otherwise. It is expected that the alignment score is short not only when the two sequences are equal but also when they contain units easy to confuse with each other. To avoid this drawback, penalty weights for the errors committed can be included in the alignment procedure. To keep the summation criterion ($\sum_i P_i = 1$), if errors are penalized, hits should be rewarded in a way to satisfy it.

The value of the score largely depend on the type of units aligned, on the length of the sequences involved and on the confusion matrix used to generate the weights. To tackle the first issue, some type of errors could be considered "less harmful" than others. Thus, the alignment procedure should be able to disregard some errors produced by "inoffensive" units such as silence. The length issue can be solved by normalizing the resulting score by the length of the input signal. The confusion matrix to be used, deserves further attention. The use of a alternative–principal confusion matrix to score the sequence alignment seems to be right choice. However any of the recognition systems produces absolute correct results. Reality is only in the actual speech transcription. We are not looking for a second opinion about some statement that might be biased (principal recognition), instead we are looking for a second opinion concerning the reality. This lead us to the need to include, somehow, the information contained in the confusion matrices of the reference (actual speech transcription) against the two kinds of recognition. One possible approach would be to use the probabilities of the reference–principal and reference–alternative matrices to generate a compound probability that, in some sense, expresses the confusion between reference and both recognizers.

If recognition events are independent, we can easily combine their probabilities:

$$P(s_j, t_k|r_i) = P(s_j|r_i)P(s_k|r_i) \qquad j \neq k \tag{2}$$

where s_j is the principal recognition unit, t_k the alternate one and r_i is the reference unit. The indexes of the sequence slots are $i, j, k = 0, 1, \ldots, N$. N is the number of phonetic units involved and can be 0 for insertions and deletions. A relevant detail should be noticed, when $j = k$, independence is not guaranteed but compound probability can be computed with:

$$P(s_j, t_j|r_i) = P(s_k|r_i) - P(t_j|r_i) + P(s_j|r_i)P(t_j|r_i) \qquad k = j. \tag{3}$$

However, this combined probability is not very useful at all. A quick glance at the results of the principal recognizer reveals that, due to its high accuracy level, its confusion matrix is rather diagonal. Therefore, it is natural to obtain zero valued compound probabilities and to cut the aligned path before arriving to the end. There is the need to avoid the zeros in the confusion matrix. To solve this question, we propose two alternative approaches to generate a compound probability: to use the maximum or the minimum of both probabilities.

Further reflection about the zero valued probabilities in the alignment process leads us to some different approach: so far we have been using the compound probabilities for two purposes: to calculate the alignment score and to define the ideal path. The ideal path defined in this way does not necessarily correspond to the ideal path that an evaluator

would build because the weights used to generate each of them are different. If we use some alignment weights similar to the ones used in evaluation to calculate the alignment score that defines the path and, by the other hand, we define a "sequence score" that results from multiplying the confusion probabilities of the units contained in the ideal path, we would be able to avoid the zeros in the confusion probabilities.

3 Opinion polling system

The opinion polling system can be implemented by submitting the same speech input to two different recognition systems and then to compare their results. Each system plays a different role: there is a principal recognizer with high accuracy level; largely equipped with capabilities to handle the vocabulary to be recognized, and with high performance phonetic units (Demiphones [4]). The recognition results from this system are the ones to be validated by the results of the alternative recognizer. The alternative system should be able to detect any sort of speech input (out of vocabulary words and noise included) and, therefore, should be equipped with unspecialized phonetic units (phonemes or discriminatively trained phonemes) and a non-restrictive (or even null) LM. The results from a system as such are not reliable at all and cannot be considered as recognition hypotheses, but are a good point of reference to compare the principal recognition with. When some utterance gives an alternative recognition result similar to the principal hypothesis, we can surely assume that the recognition of that utterance has been "clear" enough so even the alternative system could correctly recognize it. The main drawback of this scheme is that it is not that determinant when the distance between sequences is large. This could mean that the principal recognition is correct but the alternative largely incorrect. The latter let us foresee that the score sequence cannot be used as a reliable CM by itself. Nevertheless, this does not mean that this feature cannot be used as another knowledge source to build a combined CM.

To combine knowledge sources into a CM, several authors have proposed different perspectives including Bayesian classifier [5], linear discriminant analysis [6], neural networks (more specifically a multi-layer perceptron [7]) or decision trees [8]. Due to its capabilities to deal with imprecise knowledge and linguistic variables, we propose the use of a Sugeno type Fuzzy Inference System [9] as uniting tool. A Fuzzy system is able to map gradual levels of features into a confidence degree. Our FIS is based on a set of six "if ... rules" that relate the values of both features with a corresponding confidence value.

To know how the overall method should be used and and how to fix the values of its parameters in order to achieve the highest performance possible, it should be submitted to experimentation and evaluation.

4 The second opinion tested

The described system has been tested under a keyword verification task. It is, consequently, an isolated words verification system. The data base used for testing is the Spanish part of SpeechDat [10], more specifically, the city names part of SpeechDat. This is database collected through the fixed telephone network, sampled at 8 kHz and recorded under several acoustic environments. The test set to be recognized includes 414 Spanish cities names uttered by different speakers. Only the 50 % of them actually contain one of the 41 predefined keywords (207 utterances for an average of around 5 utterances for each

vocabulary word), the rest of the utterances contain one of 134 city names not related to the vocabulary ones and referred to as "out of vocabulary" (OOV) names. The task for the verifier is to validate those vocabulary names that have been correctly recognized and to reject wrong ones.

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 HMM's with quantization to the 6 (2 for the energy) closest codewords. The codebook size was 128 (32 for the differential energy). Phonetic models training was performed with a maximum-likelihood (ML) criterion with a set of 1000 phonetically rich phrases (also taken from SpeechDat). Exception made for the discriminative Phoneme set trained with a discriminative criterion.

The principal recognition system is a SCHMM based recognizer equipped with an strict LM that concatenates a set of 327 state-tied Demiphones [4] into one of the 41 city names (vocabulary keywords) to be recognized. Thus, its results always produce a vocabulary output even when OOV inputs are present. For the alternative network, two sets of 26 Phonemes plus silence, trained under a ML and under a discriminative framework, were used as phonetic units. A "grammar-free" LM allows any Phoneme string. In order to regulate the configuration of the alternative, some restrictions are added in terms of a trigram and transition penalties in the LM. Transition penalties are from two flavors: multiplicative and additive and affect transitions according to:

$$\log P_w(\lambda_i, \lambda_j) = M \times P(\lambda_i, \lambda_j) + S \tag{4}$$

where M is the multiplicative weight, S is the additive one and $P(\lambda_i, \lambda_j)$ is transition probability between the HMM's λ_i and λ_j.

A verification process has two purposes: to detect the maximum number of correct results while rejecting, at maximum, erroneous results. Two kinds of error may arise: false alarms (wrong or OOV instances taken as correct results) and false rejections (correct detection wrongly considered as OOV ones or as missed recognitions). If the verifying threshold is very permissive, it will convey a large detection rate but with the inconvenient of generating also a large number of false alarms. A less permissive threshold will reduce the number of false alarms but also will decrease the detection rate. Graphically, the relation between the detection rate and the false alarms tolerated can be expressed by means of the plots known as ROC (receiver operation curve). This plots represent our principal way to evaluate discriminative properties. Detection rate is expressed in terms of percentage, being 100 % of the vocabulary words the maximum. This upper limit is not reached in our experiments because the recognizer is not perfect. Without any false alarm rejected, it correctly recognizes 95.2 % of the vocabulary words. On the other hand, the maximum number of false alarms that the system could generate is the sum of the OOV instances (207) and the wrongly recognized vocabulary words (10) that equals 217 possible false alarms. The verifying system should be able to achieve the highest detection rate while avoiding at maximum the number of false alarms. Aside of ROC's, to evaluate our procedure, we will use a measure of the cross entropy of classification (CREP) as defined in [11]:

$$CREP = \frac{1}{N} \sum_w [\delta_w log(c_w) + (1 - \delta_w) log(1 - c_w)], \tag{5}$$

where c_w is the probability that the recognized word is correct and δ_w is 1 when recognition is correct and 0 otherwise. A lower value of CREP implies a better classification

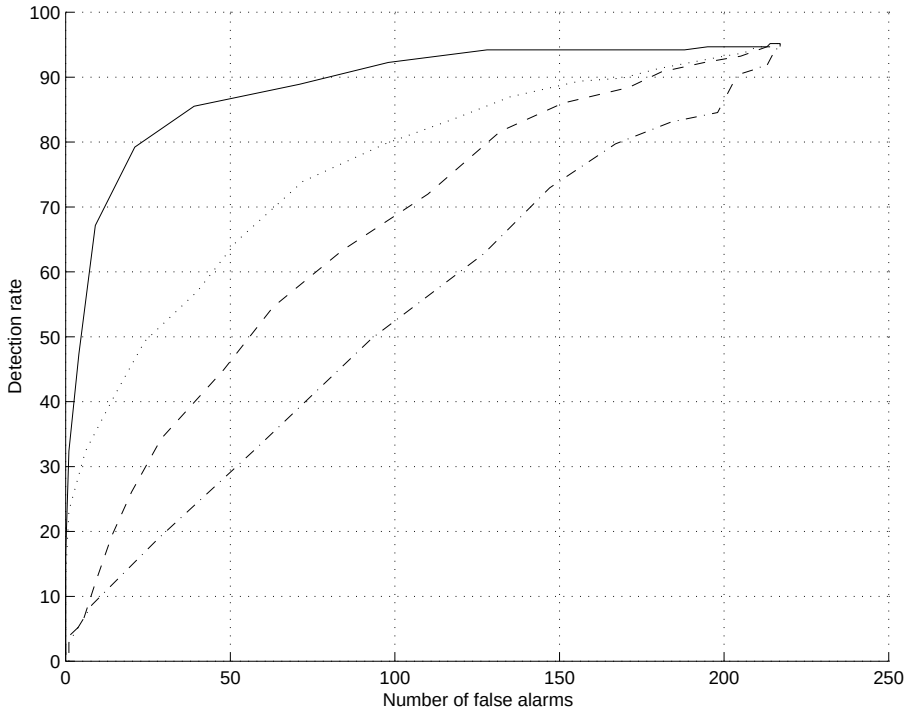

Figure 1: ROC's of (—) log-likelihood score ratio, (- -) log-likelihood of principal recognizer, (. . .) best and (. -) worst opinion polling systems

performance.

It is worth noting that the results shown here do not reflect the whole of the experiences tested. Much experimentation with the several variables that this system contains has been done but, in order to clearly express the nature of the system, only the most relevant experiments are described.

To compare the performance of the features involved in this scheme, figure 1 displays ROC's of the best and worst performing systems obtained from the opinion polling systems tried. The best one results when using discriminative Phonemes and a trigram with transitions weights $M = 1$ and $S = 3$ as acoustic and language models. For the definition of the alignment, external weights (similar to the ones used in the evaluation process from where the confusion matrices were generated) were used. A compound confusion probability for every pair of aligned phonetic units was calculated by taking the maximum of the principal–reference and alternative–reference probabilities. Silence is avoided in the alignment and a weight penalty of 0.7 was added to errors. The worst result was produced after using a simple configuration that includes ML Phonemes without LM restrictions and confusion probabilities taken from a unique alternative–reference confusion matrix used for alignment definition and scoring and no penalties added. As baseline we consider the discriminative performance of the time-normalized likelihood score resulting from the principal recognizer. Another feature, the time-normalized likelihood score ratio is calculated as the difference of the principal recognizer log-likelihood against the alternative one.

Figure 1 prompts some interesting observations: the score ratio performs much better than the rest of the systems. The opinion polling system has good discriminative behavior but is very far from the score ratio. On the other hand, the opinion polling system performs worse than the baseline when it is not properly tuned. From the latter follows

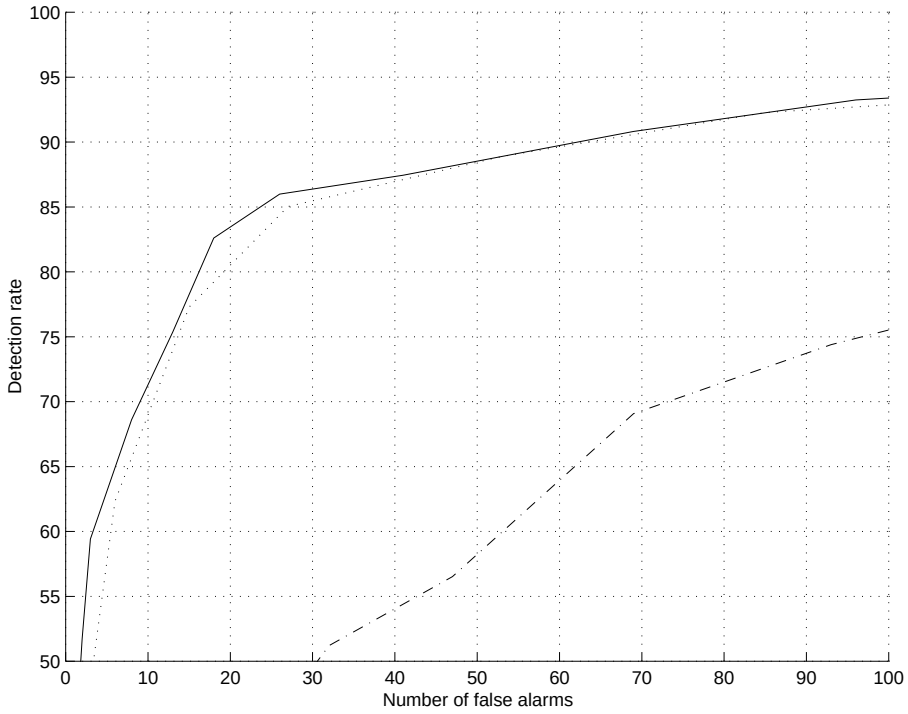

Figure 2: ROC's of (—) both features combined, (. . .) log-likelihood score ratio, (. -) opinion polling system

Table 1: Classification cross entropy of the features and of the CM

CREP seqScore	CREP scoreRatio	CREP combined
-2.054	-1.355	-0.711

that we should be very careful when using the second opinion scheme since not every configuration can be useful to discriminate. Only the score difference can be used as a reliable CM, but the purpose of knowledge sources is not to be used in isolation but rather combined. Figure 2 shows the combination of the best opinion system with the score difference by means of the previously described FIS. For a better view, the plot shows the most relevant area of the whole figure: low false alarms and high detection rate.

Results show an important improvement of the combination respect to the score ratio, mostly on the low false alarms ratio where 84 % of the correctly recognized words are detected with just 20 false alarms added. For what CREP is concerned, table 1 shows a considerable reduction of the entropy when both features are combined into a new and enhanced CM.

5 Conclusions and on-going work

The second opinion system consists in a novel approach for extracting information useful to evaluate confidence of recognition results. It only uses information from the recognition results by themselves. Therefore, it is not necessary to know the nature of the whole recognition procedure to be done. This fact eases the implementation of a posteriori CM

generators. Its configuration is a delicate issue. However, when it is properly tuned, it can represent a very useful knowledge source to build more efficient CM. Results show an important improvement of the discriminative power of combined CM compared to likelihood score ratio. Fuzzy systems represent an straightforward and effective means to compile recognition features into CM. Their versatility and capacity to deal with imprecise quantities demonstrate so. A natural extension of the present work is to apply this approach to continuous speech verification. To include a self-learning (under a back-propagation framework) FIS and to add more useful and effective features to the CM generation process represents our currently on-going work.

References

[1] R. C. Rose and D. B. Paul, "A Hidden Markov Model based keyword recognition system", in *Proceedings of 1990 ICASSP*, Albuquerque, April 1990, vol. I, pp. 129–132.

[2] S. R. Young and W. Ward, "Recognition confidence measures for spontaneous spoken dialog", in *Proceedings of EUROSPEECH'93*, Berlin, September 1993, vol. II, pp. 1177–1179.

[3] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting", in *Proceedings of 1995 ICASSP*, Detroit, April 1995, vol. I, pp. 297–300.

[4] J. B. Mariño, P. Pachès-Leal, and A. Nogueiras, "The demiphone versus the triphone in a decision-tree state tying framework", in *Proceedins of 1998 ICASSP*, Seattle, May 1998, vol. I, pp. 477–480.

[5] S. Cox and R. C. Rose, "Confidence measures for the Switchboard database", in *Proceedings of ICSLP'96*, Philadelphia, October 1996, vol. I, pp. 478–481.

[6] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition", in *Proceedings of 1997 ICASSP*, Munich, April 1997, vol. II, pp. 875–878.

[7] P. Modi and M. Rahim, "Discriminative utterance verification using multiple confidence measures", in *Proceedings of EUROSPEECH'97*, Rhodes, September 1997, vol. I, pp. 103–106.

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

[9] J. M. Mendel, "Fuzzy logic systems for engineering: a tutorial", *Proceedings of the IEEE*, vol. 83, no. 3, pp. 345–377, March 1995.

[10] A. Moreno and R. Winsky, "Spanish fixed network speech corpus", Tech. Rep., SpeechDat Project LRE-63314, 1997.

[11] M. Weintraub and F. Beaufays et al, "Neural - network based measure of confidence for word recognition", in *Proceedings of 1997 ICASSP*, Munich, April 1997, vol. II, pp. 887–890.