A Confidence Measure Invariant to Language and Grammar

Daniele Colibro, Luciano Fissore*, Claudio Vair*, Emanuele Dalmasso^, Pietro Laface^*

Loquendo, Torino, Italy*
Politecnico di Torino, Torino, Italy^
{Daniele.Colibro, Luciano.Fissore, Claudio.Vair}@loquendo.com
{Emanuele.Dalmasso, Pietro.Laface}@polito.it

Abstract

Confidence measures are necessary in all voiced activated applications to decide whether a recognized word, or a sentence, should be accepted or rejected.

A confidence measure should not only be reliable, but possibly application independent, i.e. its dynamic range should be uniform for different languages, grammars, and vocabularies. This is an important practical issue because it allows the application developers to use the same value of the threshold for different applications and to expect comparable rejection rates. This eases their task at least in the first phase of application development.

In this paper, we introduce a confidence measure that has these properties. It allows eliminating the cumbersome experimental procedure necessary to tune individually the rejection threshold for every developed recognition object.

We present the results of a set of experiments that demonstrate the “normalization” quality of our confidence measure for six different grammars in different languages.

1. Introduction

The confidence measures are used in most telephone applications to allow the dialog system to rely on the (parts of) sentences that have been reliably detected. These applications often make use of continuous speech recognition, controlled by grammars of different complexity, for carrying out their task.

In [1] we presented the results of a set of experiments aiming at assessing the quality and the limitations of different confidence measures for six different grammars.

A confidence measure should not only be reliable, but possibly application independent, i.e. its dynamic range should be uniform for different languages, grammars, and vocabularies. This is an important practical issue because it allows the application developers to use the same value of the threshold for different applications and to expect comparable rejection rates. This eases their task at least in the first phase of application development. Moreover, as pointed out in [2], it may happen that the preset rejection threshold may no longer be optimal if task adaptation is performed.

In this paper, we introduce a confidence measure that has these properties. It allows eliminating the cumbersome experimental procedure necessary to tune individually the rejection threshold for every developed recognition object.

The paper is organized as follows: Section 2 gives a short overview of the Loquendo ASR system. Section 3 details the confidence measure that was previously used in the system, while the new confidence measure is introduced in Section 4. Section 5 is devoted to the comparison of the results.

2. Loquendo ASR system overview

The Loquendo-ASR decoder uses a hybrid HMM-ANN model where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops. The models are based on a set of vocabulary and gender independent units including stationary context-independent phones and diphone-transition coarticulation models. A Multilayer Perceptron estimates the posterior probability of each unit state, given an acoustic feature vector. The HMM transition probabilities are uniform and fixed [3].

The confidence measures used in this work are based only on the posterior probabilities generated by the decoder.

3. Original confidence measure

Confidence measures based on local phone posterior probability estimates generated by a hybrid HMM/ANN model have been proposed in [4,5]. To account for the raw acoustic information associated to each frame, the best score has been proposed as a measure of the matching between the data and the model [6]. In this approach, each utterance frame is scored against every output distribution in their HMMs to find the best score, independent of any information given by the sequence of phonetic units or words.

Building on these ideas, we have proposed in [1] as a confidence measure the Acoustic Log Likelihood Ratio defined as:

$$ALLR = \frac{\sum_{t=1}^{T} \log P(s_{i^*} \mid o_t)}{\sum_{t=1}^{T} \max_{1 \le j \le S} \log P(s_j \mid o_t)} \qquad (1)$$

where S is the set of output states of the ANN model, o_t is the *t-th* acoustic observation vector, and s_{i^*} is the sequence of states - indexed by i^* - produced by the Viterbi alignment of an utterance of T observation frames.

ALLR is the ratio between the sum of the frame scores constrained by the model of word *w* and the free score, given by the sum of the a posteriori log probability of the best matching state for each frame. This measure is easily obtained in a hybrid HMM/NN model because all the posterior probabilities are computed in parallel by the NN. The value of *ALLR* ranges from 0 to 1, and its maximum is reached when the free and the constrained scores are the same for each frame, denoting an optimal acoustic matching according to the model. Low values of *ALLR* are, instead, good indicators of acoustic mismatch. According to the observation interval, the *ALLR* confidence measure computes the reliability of an acoustic-phonetic unit, of a hypothesized phone, of a word, or even – excluding silence intervals – of a sentence.

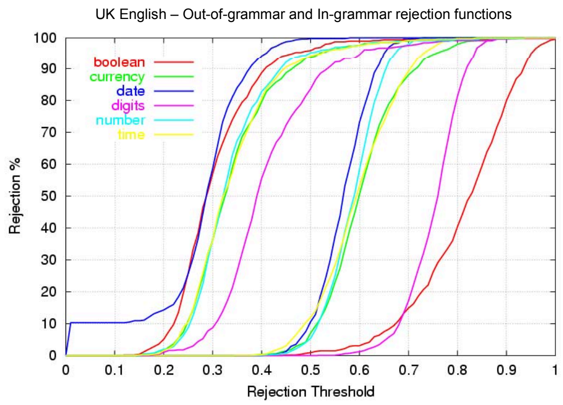

Figure 1: Rejection functions for 6 UK-English grammars

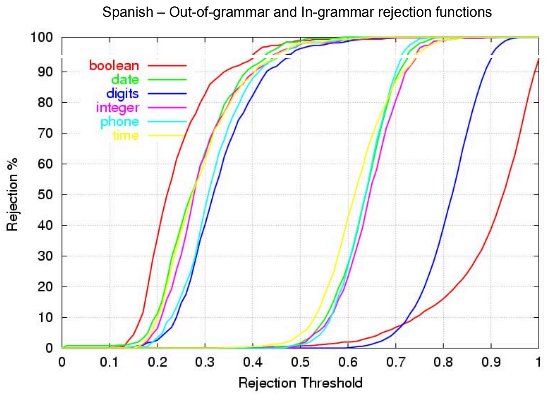

Figure 2: Rejection functions for 6 built-in Spanish grammars

For the rejection of out of grammar utterances, the confidence measures are combined in different ways to obtain a confidence measure at the sentence level [7]. In our experiments, for all the grammars, the best combination of the word level confidence measures for detecting out of grammar sentences is the average of the confidence scores of the words in the sentence [1], but similar results are obtained using (1) for the whole sentence frames, excluding the silence intervals.

Figures 1 and 2 show the rejection functions for six built-in UK-English and Spanish grammars respectively. The figures plot the rejection rates obtained as a function of a given confidence threshold. The family of curves on the right of the figures refers to in-grammar utterances, while the family of curves on the left refers to out of domain utterances.

The analysis of these results, similar to those referring to other languages, clearly shows a high heterogeneous behavior of the rejection curves: the same confidence threshold value produces considerably different rejection rates for different grammars and languages. In particular, the rejection curves for in-grammar utterances show large spreading, while we are mainly interested in controlling the false rejection rate for in-grammar utterances.

To reduce the effects of this variability, the built-in grammars of the Loquendo ASR system were released with specific object and language dependent thresholds in order to obtain, for the same threshold, comparable rejection behaviors.

This compensation was, of course, impossible for grammars defined by the users. The rejection threshold of every developed recognition object had to be tuned individually.

4. New confidence measure

Since this approach was only corrective, and moreover inappropriate for user-defined objects, a new confidence measure invariant to language and grammar has been devised. As it has been pointed out in the previous Section, the confidence measure of (1) produces rejection functions that lack in stability and homogeneity. This is mainly due to the variability of the a posteriori probability distributions of the acoustic-phonetic units. The a posteriori probability distribution of an acoustic-phonetic unit is affected by several factors such as, for example, the acoustic characteristics of the unit, its occurrence in the words of a given language, its confusability with other units, the structure of its model, and the amount of available training data. Since the confidence measure is derived from the a posteriori probabilities of the units, the variability of the latter produces instability and lack of homogeneity in the confidence itself.

4.1. Normalized differential confidence measure

The first step made toward reducing this variability has been to define a confidence measure that is a minor variant of (1):

$$DC = \frac{1}{T}\sum_{t=1}^{T}\log P(s_{i^*} \mid o_t) - \frac{1}{T}\sum_{t=1}^{T}\log\left[\max_{1\le j\le S} P(s_j \mid o_t)\right] \qquad (2)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\log\left[\frac{P(s_{i^*} \mid o_t)}{\max\limits_{1\le j\le S} P(s_j \mid o_t)}\right]$$

This confidence measure can be interpreted as the average of the confidences computed frame by frame. It produces negative values, and zero that represents the maximum reliability. Experimental results show that the quality of this measure and the behavior of its rejection functions are equivalent to the *ALLR* confidence of (1). This measure is interesting because it is possible to weight the terms of the sum as a function of the state. Moreover, it is possible to select the contributions to the average, excluding, for example, the silence frames.

To introduce the second step we must recall that in a hybrid HMM-ANN model, the output layer produces for each acoustic state s_i an activity value n_i. The activity values are filtered by a sigmoid function (3) to obtain for each state s_i its a posteriori probability:

$$P(s_i \mid o_t) = \frac{e^{n_i(o_t)}}{\sum\limits_{j=1}^{S} e^{n_j(o_t)}} \qquad (3)$$

The contribution of a frame to the new confidence is, thus:

$$C(i^*,t) = \log\frac{P(s_{i^*} \mid o_t)}{\max\limits_{1\le j\le S} P(s_j \mid o_t)} = \log\frac{e^{n_{i^*}(o_t)}}{e^{\max\limits_{1\le j\le S} n_j(o_t)}} \qquad (4)$$

$$= n_{i^*}(o_t) - \max_{1\le j\le S} n_j(o_t)$$

We define a new differential confidence measure as:

$$DC = \frac{1}{T}\sum_{t=1}^{T}\left[C_p(i^*,t)\right] = \frac{1}{T}\sum_{t=1}^{T}\left[n_{i^*}(o_t) - \max_{j\in phones} n_j(o_t)\right] \qquad (5)$$

where the best matching state for each frame is selected among the *stationary context-independent phones* only. The selection

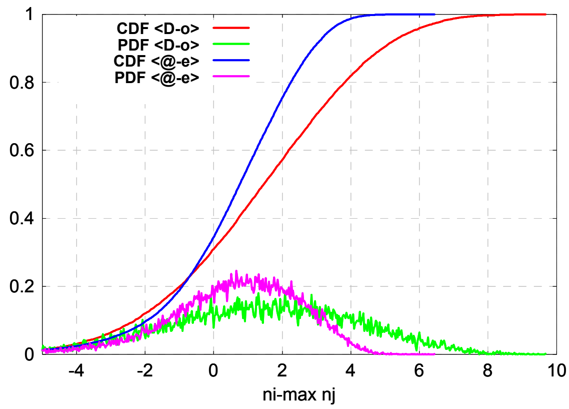

Figure 3: PDFs and CDFs of two transition unit states of the Spanish language.

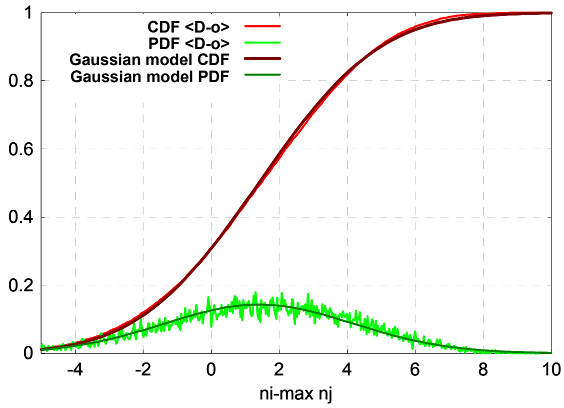

Figure 4: PDF, CDF, and estimated Gaussian model of a transition unit state of the Spanish language.

is limited to phones because, for the sake of efficiency, our ANN computes the probabilities of all phones, but only the probabilities of the transition units appearing in the application grammar vocabularies. The DC measure, therefore, is the average of the difference between the network output - $n_i*(o_t)$ - obtained using grammatical and lexical constraints, and the *best phone* output - $max_j\ n_j(o_t)$ - obtained relaxing the constraints.

4.2. Differential confidence normalization

To reduce the variability of the a posteriori probabilities of the acoustic states, statistics of the distribution of the differences n_{i^*} - $max_j\ n_j$, for each state s_i have been collected. Considering these values instances of a random variable ξ_i, the corresponding probability density functions (PDF) f_ξ , and the cumulative distributions (CDF) F_ξ have been computed. These functions account for the characteristics of each acoustic state and reveal the differences between states.

Figure 3 shows the probability density and the cumulative distributions functions computed for a state of two transition units of the Spanish language. The symbol <@> in the figure refers to the silence unit.

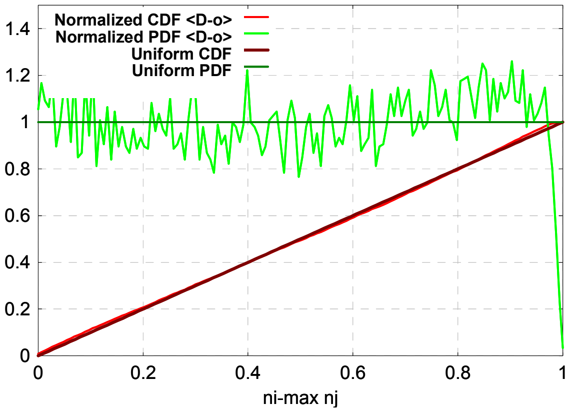

Figure 5: PDFs and CDFs normalized by means of F_ξ, and the corresponding uniform distributions.

To make homogeneous the behavior of all the states we take advantage of a cumulative distribution function property: F_{ξ_i} applied to the random variable ξ_i from which it has been estimated produces a new random variable estimated η=$F_{\xi_i}(x)$ uniformly distributed in the range [0-1]. The cumulative distributions have been obtained, through forced alignments of the training data, estimating the mean and the variance of the contributions $C_p\ (i^*,t)$ for every state assuming that the contributions have a Gaussian distribution.

Figure 4 shows the estimated Gaussian distributions for a two transition unit state of the previous example.

The results shown in Figure 5 are obtained, instead, applying the probability density function to the corresponding random variables, and re-estimating their distribution.

Since the resulting distributions are good approximations of a uniform distribution, the values of the normalized contributions $F_{\xi_i}\ (C_p(i^*,t))$ are good candidates for a more homogeneous confidence measure.

Thus, the normalized differential confidence is computed as:

$$NDC = \frac{1}{T}\sum_{t=1}^{T} F_{\xi_i}\left[n_{i^*}(o_t) - \max_{j\in phones} n_j(o_t)\right] \qquad (6)$$

Since the normalized differential confidence *NDC* is the average of random variables uniformly distributed in the range [0-1], it assumes values in the same range, and thus can be directly used as a confidence measure. It is, furthermore, worth recalling that, according to the central limit theorem, the distribution of the NDC values is not uniform, but Gaussian with mean 0.5. As shown in Figure 6, the curves that represent the cumulative distributions of the NDC variable are indeed fairly Gaussian, although their mean value is slightly greater than 0.5. This is reasonable because the statistical independence hypotheses and of homogeneity of the distributions of the terms contributing to the normalized differential confidence (6) are not exactly verified.

To better exploit the interval [0-1], and to further reduce the heterogeneity among different languages, a linear compensation

$$y = \alpha * NDC + \beta \qquad (7)$$

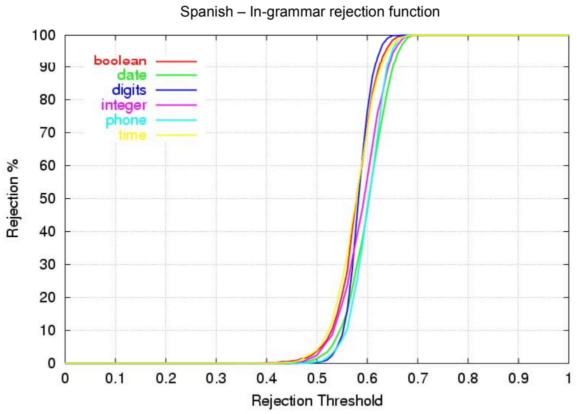

Figure 6: Rejection functions for the six built-in Spanish grammars using the normalized differential confidence.

is performed on *NDC* to obtain an average rejection level of 5% for correct recognitions using a confidence value of 0.65 and of 95% of rejection for a confidence value of 0.90.

The compensation factors are the same for all the objects of a given language, and also similar among languages.

Figures 7 and 8 show examples of rejection functions obtained, for the same six built-in UK-English and Spanish grammars referred to in Figures 1 and 2, using the new normalized differential confidence measure.

Comparing the functions of Figures 1 and 2, obtained using the old confidence measures, with the ones of Figures 7 and 8, one can immediately appreciate how the family of curves related to in-grammar utterances has been clustered toward similar distributions for the different grammars and languages, allowing to preset uniform rejection thresholds.

Similar results have been obtained for the built-in grammars of the 15 languages released with the Loquendo ASR.

It is also worth noting that we could normalize the contributions $C_p(i^*,t)$ to the normal distribution $N(0,1)$ simply through mean and variance compensation. This procedure would give, after an appropriate rescaling, similar results with respect to the approach using the cumulative functions and the uniform distributions. The advantage of using the CDF is that it could allow better numerical fitting to the original distributions. Moreover, the CDFs can be easily tabulated for a very fast lookup access. That is why this approach has been followed in this work.

5. Conclusions[1]

We presented an approach based on the computation of a posteriori probabilities that allows computing a confidence measure homogeneous in terms of rejection capabilities, and invariant to the recognition constraints such as the grammars, the recognition vocabularies, the language, or even the complexity or accuracy of the acoustic models. Further work will focus to reliably identify noise regions in the sentences. Since in these regions both the constrained and unconstrained probabilities are similar - and good - their contributions to the total confidence introduce a bias that must be eliminated.

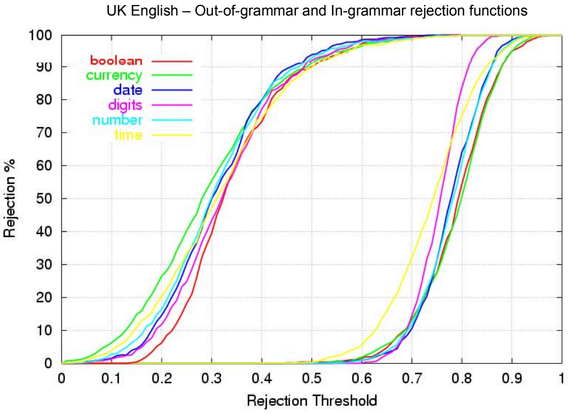

Figure 7: Rejection functions for the six UK-English grammars using the rescaled normalized confidence.

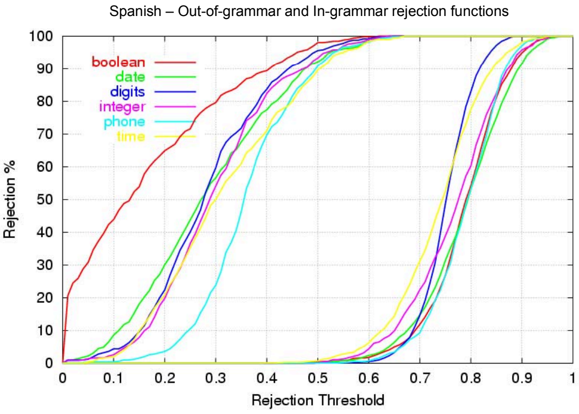

Figure 8: Rejection functions for six built-in Spanish grammars using the rescaled normalized confidence.

6. References

[1] M. Andorno, P. Laface, R. Gemello, “Experiments in Confidence Scoring for Word and Sentence Verification,” ICSLP-2002, pp. 1377-1380, 2002.

[2] A. Sankar, A. Kannan, “A comprehensive study of task-specific adaptation of speech recognition models”, Speech Communications, Vol. 42, pp. 125-139, 2004.

[3] D. Albesano, R. Gemello, F. Mana, “Hybrid HMM-NN Modelling of Stationary-Transitional Units for Continuous Speech Recognition”, Int. Conf. On Neural Information Processing, pp. 1112–1115, 1997.

[4] G. Williams, S. Renals, “Confidence Measures from Local Posterior Probability Estimates”, Computer Speech and Language, Vol. 13, pp. 395–411, 1999.

[5] G. Bernardis, H. Bourlard, “Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems”, ICSLP 1998, pp. 775–778, 1998.

[6] L. Gillick, Y. Ito, J. Young, “A Probabilistic Approach to Confidence Estimation and Evaluation”, ICASSP 1997, pp. 879–882, 1997.

[7] B. Souvignier, A. Wendemuth, “Combination of Confidence Measures for Phrases”, *Proc. ASRU 1999 Workshop*, Keystone, USA, pp. 217-220, 1999.

[1] This work was partially supported by the EU FP-6 IST Project DIVINES – Diagnostic and Intrinsic Variabilities in Natural Speech