On Combining Confidence Measures for Improved Rejection of Incorrect Data

Delphine Charlet, Guy Mercier & Denis Jouvet

France Télécom R&D - DIH/IPS
2 av. Pierre Marzin, 22307 Lannion, France
(delphine.charlet,guy.mercier,denis.jouvet)@francetelecom.com

Abstract

In this paper, techniques for combining confidence measures are proposed and evaluated. Confidence measures are useful for rejecting incorrect data, which is an important issue in speech recognition based interactive systems. Many ways of computing individual confidence measures have already been investigated. A detailed analysis of various confidence measures shows that they behave differently for what concerns rejection of incorrect data on various field data subsets (substitution errors, out-of-vocabulary data & noise tokens) collected from a vocal directory task. Two combination techniques are then presented. One combines confidence measures by means of a neural network and the other through logistic regression. Evaluations shows that both combination techniques are efficient, and both take the best of the various individual confidence measures involved on each data subset.

1. Introduction

A reliable estimate of the confidence with which a user utterance has been recognised is essential for user-friendly human-machines dialogs, and is also useful for unsupervised adaptation. Confidence measures provide an estimate of the reliability of the speech recogniser’s output. Its comparison to a decision threshold provides a mechanism for accepting or rejecting the speech recogniser’s answer. Hence this allows to reject incorrect data.

Many confidence measures and rejection strategies have already been investigated. A currently used rejection strategy relies on the introduction of a garbage model, which amounts to computing a likelihood ratio between the word or sentence model and a background model (the garbage model). Other confidence measures derived from word graphs are often used in continuous speech recognition. The likelihood ratio between the best and second-best decoding answers also provides an efficient confidence measure [1, 2]. Besides, confidence measures based on hypothesis testing approach [3] have also been investigated. Such approaches rely on a double modeling; one is associated to correct events, and the other to incorrect events. They are used to compute a likelihood ratio test statistic. This has been applied to various parameters ranging from segmental prosodic and phonetic features [4, 5] to frame features.

A detailed analysis of several confidence measures on various field data subsets (substitution errors, out-of-vocabulary data & noise tokens) collected from a vocal directory task showed that they behave differently on the different subsets. Hence, it would be useful to combine them in such a way as to get the best of each individual measure on each data subset. In order to achieve this goal 2 combination techniques are investigated and evaluated. One is based on a neural network approach and the other relies on a logistic regression function.

The paper is organized as follows. Section 2 describes the individual confidence measures. Section 3 presents the 2 combination techniques. Section 4 introduces the experimental setup. Finally results are presented and discussed in section 5.

2. Confidence measures

Several confidence measures are investigated. A few of them are based on segmental phonetic features, which are estimated by means of fuzzy neural networks. The last one is derived from the likelihood ratio between the 2-best decoding answers.

2.1. Segmental phonetic-based confidence measures

This approach assumes that each segment composing a recognition hypothesis $W$ (a word or a sequence of words) is characterized by a set of phonetic features like voiced/unvoiced, vocalic/consonant, nasal/oral, and so on. A neural network (NN) trained to recognize a specific phonetic feature $j$ is able to compute a phonetic NN segmental feature noted $x^j$, for each labeled phonetic segment $\phi_i$ given by the Viterbi alignment associated to an hypothesis. So given an hypothesis $X$, recognized as $W$, it is then possible to compute the following log likelihood ratio test statistic $LLR(X \mid W)$:

$$LLR(X \mid W) = \sum_{i=1}^{\text{word}} \log \frac{P(x^j \mid M_{\phi_i})}{P(x^j \mid M_{W})}$$

(1)

where $M_{\phi}$ and $M_{W}$ are respectively the model and anti-model associated to the phoneme $\phi_i$, and $\text{nsseg}$ is the number of segments composing the hypothesis $X$.

2.1.1. Phonetic feature estimation by neural network

The neural network architecture (Figure 1) used to decide whether a phonetic feature is present or not is derived from a neural-fuzzy network designed by Glorennec [6].

For each input frame of a segment, the first layer is made up of 27 cells corresponding to 8 Mel frequency cepstral...
coefficients and the logarithm of the frame energy plus their first and second order derivatives; the role of the second layer is to compute the membership values associated with each input and with each fuzzy subset; this is done by means of gaussian functions (2 gaussian functions for each input).

$$\sum_{i} x_{i} \beta_{i} + \sum_{i} x_{i} \beta_{i}$$

where the weight $\beta$ depends on the number of samples available for estimating the context-dependent model $M_{\text{CD}}^c$.

Obvious correct events are associated to segments belonging to correct alignments (i.e. correct recognition). But for what concerns incorrect events, not all the segments of incorrect alignments are incorrect. A recognition error may occur even if some segments of the alignment are correct (right phoneme at the right place). Hence, for training anti-models, each incorrect alignment is compared to the associated correct alignment in order to determine which segments are correct (same phoneme at almost the same place) and which are not. Moreover as the NN are trained for computing phonetic features, a segment is considered "incorrect" (i.e. used for training anti-model) only if the associated phoneme has a phonetic feature different from the phonetic feature of the reference correct segment.

2.2. Confidence measure based on N-best decoding

The confidence measure used in this work is the normalized log-likelihood ratio between the first and second best hypotheses in a N-best decoding approach. The log-likelihood ratio is normalized by the length of the utterance (in frames). This confidence measure is a classical one when N-best decoding is available, and has proven to be quite effective [1, 2]. For a given utterance $X$ of T frames for which the first hypothesis is $W_i^1$ and the second one is $W_i^2$, the confidence measure is given by:

$$\frac{LLR_i(X | W_i^1)}{T} = \frac{1}{T} \log \left( \frac{P(X | W_i^1)}{P(X | W_i^2)} \right)$$

3. Combining confidence measures

Two techniques for combining confidence measures are proposed here. The first one relies on a logistic regression function whereas the second one uses a neural network.

3.1. Logistic regression based combination

Logistic regression is a formalism that enables to fuse predictors and to give a response mapped to a probability. It is based on the assumption that the log-likelihood ratio for a set of predictors (here confidence measures $cm_i$) can be estimated by a linear combination. For instance, for the fusion of $N$ predictors $\{cm_1, \ldots, cm_N\}$:

$$\log \frac{p(cm_1, \ldots, cm_N \mid \text{correct})}{p(cm_1, \ldots, cm_N \mid \text{incorrect})} = b_1 + b_2 cm_1 + \ldots + b_N cm_N$$

hence:

$$P(\text{correct} \mid cm_1, \ldots, cm_N) = \frac{1}{1 + \exp^{- (a_0 + a_1 cm_1 + \ldots + a_N cm_N)}}$$

where $A = \{a_0, a_1, \ldots, a_N\}$ is the vector of coefficients and $CM = \{cm_1, \ldots, cm_N\}$ the vector of predictors.

Thus, logistic regression enables the estimation of the posterior probability that the response of the classifier is correct, given all the confidence measures.
The coefficients $A$ of the logistic regression function are estimated in order to maximize the development set likelihood:

$$L = \sum_{i=1}^{n} c_i \log p_i + (1 - c_i) \log(1 - p_i)$$  \hspace{1cm} (7)

where for each sample $i$:

- $c_i = 1$ if test $i$ is correct, $c_i = 0$ otherwise,
- $p_i$ is the posterior probability, given by Eq. 6, that the test $i$ is correct, given the vector of confidence measures $CM_i$ associated to test $i$.

As the equations ($\frac{dL}{dA} = 0$) for which the development set likelihood is maximum have no known analytic solution, they are solved in an iterative way using the Newton-Raphson method.

### 3.2. Neural network based combination

For a given phonetic feature $j$, and a recognition answer $W$, the log likelihood ratio $LLR_j(X | W)$ given by Eq. 1 is combined with the HMM log likelihood $LL Holl(X | W)$ in order to obtain the global score $SC_j(X | W)$:

$$SC_j(X | W) = \alpha LL Holl(X | W) + (1 - \alpha) LLR_j(X | W)$$  \hspace{1cm} (8)

where the interpolation coefficient $\alpha$ is optimized on the training set in order to minimize the recognition error rate.

These global scores $SC_j(X | W)$ are then merged together to derive a combined confidence measure:

$$SC_{CN}(X | W) = SC_1(X | W), \ldots, SC_j(X | W), \ldots, SC_n(X | W)$$  \hspace{1cm} (9)

where $n$ is the number of confidence measures. This function is approximated by means of a multi-layered perceptron. The input layer of this perceptron is composed of $n + 1$ cells, with the $n$ first cells being filled with the scores $SC_j(X | W)$ and the last one with the constant 1. This layer is totally connected to the second hidden layer composed of $Q$ active cells (in our experiments $Q = 6$) and the last layer is reduced to 1 cell. The final score $SC_{CN}(X | W)$ is obtained as the output of this cell.

The weights (parameters) of the perceptron are estimated on a training set by means of a classical back-propagation gradient algorithm. The target set is set to 1 when the hypothesis is correct and to 0 in the other case.

When the 2-best solutions log-likelihood ratio confidence measure $LLR_2(X | W_i)$ is used in the above combination, it is treated as one of the global score $SC_i(X | W)$.

### 4. Experimental Setup

#### 4.1. Field database

The speech database was collected from a telephone directory task in operation. Users were allowed to pronounce either the surname alone, or the first name followed by the surname. The database collected from several thousands calls over several months is used here. Part of the data (corpus *Ctrain* - 15244 tokens) is used to estimate the parameters of the models (acoustic HMM models as well as models and anti-models for computing segmental phonetic based confidence measures - cf. § 2.1.2). The remaining part (corpus *Ctest* - 6353 tokens) is used to evaluate the speech recognition performances. The test vocabulary is made of 1587 words, and is not exactly the same as the training vocabulary.

Each corpus is divided into 4 subsets. The first one, called "Data in Directory", contains only utterances that matches a directory entry. The second one, called "Data out of Directory", is also made of proper names but either the first name or the surname is not in the speech recognition vocabulary, or the sequence (first name followed by surname) does not match any directory entry. The third subset, called "Out-of Vocabulary Data", contains speech utterances which are not proper names (e.g. hesitations, incomplete entries, comments, ...). Finally, the last subset, called "Noise Tokens", is made of tokens, which do not contain speech, but were energetic enough to trigger the endpoint detector.

#### 4.2. Recognition system overview and training

The speaker-independent speech recognition system is HMM-based and relies on continuous densities. Mel frequency cepstral coefficients are computed every 16 ms, as well as their first and second order derivatives estimated over 5 frames windows. A context-dependent modeling of the phoneme units is used. The acoustic models were first trained on a large vocabulary independent database (flexible modeling), and a garbage model is also available in the modeling.

The parameters of the neural networks used to extract the segmental phonetic features were also trained on a vocabulary independent database, which was automatically segmented and labeled. The generic HMM acoustic models were adapted to the field data conditions using the *Ctrain* corpus. This was performed through an incremental EM adaptation algorithm [7]. These adapted models will serve as a reference in the following experiments.

The discrete distributions for the models and anti-models were also estimated on this corpus $Ctrain$. The parameters of the perceptron used to merge the various confidence measures were also estimated on this training corpus.

### 5. Results and discussion

Recognition error rates are reported in Figure 2 for the various data subsets. The horizontal axis always reports the false rejection rate on "Data in Directory". The vertical axis corresponds to the substitution rate for "Data in Directory" subset, and to the false alarm rate for the other subsets ("Data out of Directory", "Out of Vocabulary Data" and "Noise Tokens"). Logarithmic scales are used on each axis.

The reference curve, thick line, corresponds to the use of the garbage model alone. The curve was obtained by varying the decision threshold on the log likelihood ratio between the word or sentence model and the background (garbage) model.

When confidence measures are also used (other curves), the garbage model decision threshold is set low (hence very small false rejection rate) and applied first. Then the confidence measure based decision test is applied to every utterance that was not rejected by the garbage model based test. Curves are obtained by varying the confidence measure based decision threshold.
5.1. Comparison of individual confidence measures

Several individual confidence measures have been evaluated. One is based on the likelihood ratio between the 2 best decoding answers (cf. § 2.2). The 3 other rely on phonetic segmental features computed by neural network (cf. § 2.1.1).

It clearly appears from Figure 2, that the 2-best likelihood ratio (dotted line with circles) provides the best rejection performances for what concern Data in Directory substitution errors, and Data Out of Directory false alarms. However this measure does not bring any improvement for rejecting false alarms on Out of Vocabulary Data and Noise Tokens. On the opposite, the phonetic segmental feature based confidence measures (dotted lines with diamonds for voiced/unvoiced, squares for vocalic/consonant & triangles for speech/noise) provides better results for rejecting false alarms on Out of Vocabulary Data and Noise Tokens, but are not as efficient on the other subsets.

5.2. Combination of confidence measures

The 2 solid lines with circles and squares in Figure 2 shows the results achieved by combining the various individual confidence measures. They provide very similar results, and both combination techniques allow getting almost the best performances on each individual data subset.

6. Conclusions

In this paper several confidence measures have been compared on field data collected from a vocal directory service in operation. To better analyze the confidence measure performances, the field data was divided in 4 subsets (data in directory, out of directory, out of vocabulary and noise tokens). The different behavior of the individual confidence measures on the various data subsets lead to the idea of combining them. In order to do that, two techniques for combining confidence measures have been proposed and compared. One is based on logistic regression and the other on neural network. Both techniques prove to be very efficient and take the best of the various individual confidence measures involved on each data subset.

7. References