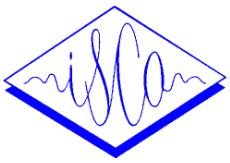

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# Confidence Measures Based on the K-nn Probability Estimator

*Fabrice Lefèvre*

Laboratoire d'Informatique d'Avignon – Université d'Avignon et des Pays de Vaucluse
BP 1228 - 84911 Avignon Cedex 09 - France

Fabrice.Lefevre@lia.univ-avignon.fr - http://www.lia.univ-avignon.fr/

## ABSTRACT

In this paper, the use of the probabilities produced by a K-nearest neighbours (K-nn) estimator as confidence measure is investigated in an hypothesis verification post-processing scheme. The objective is to classify as correct or incorrect the outputs of a Gaussian mixture model (GMM) /HMM speech recognition system. Four confidence measures based on the K-nn probability estimator are introduced. Preliminary experiments are reported and discussed on the TIMIT database.

## 1. INTRODUCTION

Confidence measure for speech recognition has become a major research issue with the development of spoken dialogue systems. The confidence measures are mostly used to help word-spotting in spontaneous speech and to provide a basis for the rejection of out-of-vocabulary words. However, many others ASR applications could also benefit from the information brings by a confidence measure. For instance, text dependent speaker recognition could put more emphasis on high level confidence segments; also unsupervised adaptation algorithms could be applied only when the confidence is high enough for the collected data.

Several methods have been yet introduced in the field of confidence measure. Although no standard classification of these methods can already be established, few categories can be distinguished. From our point of view a major distinction can be made between the confidence measures based directly upon the information derived from the decoding process and the confidence measures based on new knowledge sources complementary of those used during the decoding process. Examples of the first approach can be found in [1, 2] whereas [3] privileges the second approach.

In this paper, new confidence measures based on the K nearest neighbours (K-nn) probability estimator are investigated. One major advantage in the use of this estimator for confidence measure lies in its ability to produce *a posteriori* probabilities which represent a convenient way of assessing the trust one can give in a recognition decision.

The present study follows a previous work in which we have developed a hybrid K-nn/HMM speech recognition system [4, 5]. A complete protocol has been proposed for the training and the decoding with such a system. The baseline results were encouraging. But the state-of-the-art techniques, such as delta coefficients or contextual modelling, could not be applied advantageously in the K-nn/HMM system. The experiments have put on forth important differences of behaviour between a sate-of-the-art GMM/HMM system and the K-nn/HMM system. As a matter of fact, even when both systems obtain equivalent overall results, the alignment of their output transcriptions shows a 30% error rate. Besides it has been shown that the K-nn estimator outperforms the GMM estimator in identification tasks. Therefore, the K-nn estimator can be of practical interest to measure the reliability of a transcription obtained from a GMM/HMM system.

In this paper, it will be focussed on the use of confidence measure for hypothesis verification. Hypothesis testing and the role of confidence measure as a test statistic are described in section 3. Confidence measures are presented in more details in section 4. A set of acoustic confidence measures based on the local posterior probability estimates produced by the K-nn is introduced. These confidence measures are thereafter evaluated on the TIMIT database at the phone level by comparison with the GMM log likelihood ratio technique (section 5). First of all, the K-nn probability estimator principles and assets are recalled.

## 2. K-NN PROBABILITY ESTIMATOR

The K-nn technique is mainly known for its derived decision rule (based on a majority vote). In this study, the K-nn technique is used to provide probability estimates. Two types of probability can be computed considering a sample vector $x$ and a class $C$:

- Posterior Probabilities: $\tilde{P}(C/x) = \frac{k_C}{K}$
- Likelihoods: $\tilde{P}(x/C) = \frac{k_C}{n_C}$

with $k_c$ the number of reference samples associated with the class $C$ among the $K$ and $n_c$ the total number of reference samples associated with the class C. The K-nn probability estimator main assets are:

- it is non-parametric so it could fit any real data distributions provided that enough reference data are available ;
- the decision rule error is bounded by the optimal Bayes rule error by [6]:

$$Error_{K-ppv} < \left[1 + \frac{2}{\sqrt{K\pi/2}}\right] Error_{opt} \qquad (1)$$

- this upper bound decreases when K increases ;
- the K-nn estimator provides discriminative posterior probabilities, normalised as all the classes are considered for the estimation ;
- finally, no training phase is needed before its use.

## 3. HYPOTHESIS TESTING

A hypothesis test is a procedure which aims at deciding either to accept or to reject a predefined hypothesis (the so-called *null hypothesis*). In a one-tailed test, this decision is taken according to the value of a test statistic: below a certain threshold, the hypothesis is rejected. The performance of a hypothesis test is then represented through a 2x2 confusion matrix: -two lines for the *state of nature* of the test (its truth or its falsity), -two columns for the *decision* (accepted or rejected). Two cells of this matrix correspond to errors: -rejecting a true candidate (*Type I* or *False Rejection*), -accepting a false candidate (*Type II* or *False Alarm*). A number of statistics can be computed from a 2x2 matrix and used to evaluate the performance of the test statistic at a given operating point (fixed by the threshold value). One of these is the *unconditional error rate* (UER) of the test, which involves the sum of the number of *Type I* and *Type II* errors over the total number of hypotheses:

$$UER = \frac{N(reject, true) + N(accept, false)}{N} \quad (2)$$

The experiments assessments will be made through two curves parameterised by the decision threshold:

- the *Receiver Operator Characteristic* (ROC) is a plot of the *Type II* error rate against the Detection Rate[1];
- the *Detection Error Trade-off* (DET) is a plot of the *Type I* error rate against the *Type II* error rate.

In the context of a speech recogniser's outputs verification, the *null hypothesis* consists in the acceptance of the output. In order to carry out such a test, a test statistic should be defined. The confidence measure for a decoding hypothesis will be used as the required test statistic.

## 4. CONFIDENCE MEASURES

A confidence measure is commonly defined as a statistic upon the level of matching between a model and the data. In the case of speech recognition, two models can be considered: the acoustic model and the language model. In the further, only confidence measures derived from the acoustic model will be considered. Such purely acoustic confidence measures will be of great interest in tasks where either no good language model is enable or a great discrepancy exits between the expected and observed data.

### *4.1. Likelihood Ratios*

To measure how well a model matches the data, it is appealing to use straightforwardly the likelihood of the data upon the model used during the decoding stage.

The output likelihood of a GMM/HMM is not an appropriate statistic to be used as a confidence measure: it is relative to the prior probability of the acoustic and therefore it is not comparable across utterances. A common answer to this issue has been the introduction of anti- (or filler or garbage) models. Loosely speaking, an anti-model is composed of a set of probability distributions of all the events not associated to the target model [7, 8]. The hypothesis test is then based on a log likelihood ratio (*LLR*) which is computed for the vector-frame $x_t$ for the phonetic class *i* as:

$$LLR_i(t) = \log \frac{p(x_t|\Lambda_i)}{p(x_t|\Lambda_{\hat{i}})} \quad (3)$$

where $\Lambda_i$ is the model for class *i* and $\Lambda_{\hat{i}}$ its anti-model. A usual approach for training the anti-models is to use tokens that caused errors. Two kind of errors have generally to be considered: substitutions and noise. By noise it is meant every events which should not occur in the application. As in the context of TIMIT none noise tokens are considered, the anti-model that will be considered in the following experiments is a composition of all the GMM of the system, excepted those of the target model. Only the most significant GMM likelihood is retained (*max* operator).

At the phone level, the confidence measure is obtained by a sum of the *LLR* over the frames associated to the target model during the decoding process:

$$LLR_i(t_s, t_e) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} LLR_i(t) \quad (4)$$

where $LLR_i(t_s,t_e)$ is the confidence value for the phone *i* which has been output by the recogniser between the frames $t_s$ and $t_e$ and $LLR_i(t)$ is the frame log likelihood ratio computed with the target and anti- models as described in (3).

Although, as recalled above, one major asset of the K-nn estimator is to produce posteriors, a K-nn log likelihood ratio confidence measure (*K-nn/LLR*) will also be evaluated for the purpose of comparison with the *GMM/LLR* confidence measure.

### *4.2. Posteriors Based Confidence Measures*

From the local frame posteriors, two confidence measures are obtained depending on how they are combined at the phone level:

- Log Product Posterior confidence measure (*LPP*) which is also the log of the geometric mean:

$$LPP_i(t_s, t_e) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log p(i|x_t) \quad (5)$$

- Mean Posterior confidence measure (*MP*):

$$MP_i(t_s, t_e) = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} p(i|x_t) \quad (6)$$

[1]Actually a ROC-like will be used plotting the *UER* against the Hypothesis Rejection (*HR*) level.

Is should be noted that the time-normalisation has been applied on all the measures as several previous studies have clearly shown its good influence (see for instance [1]).

## 5. EXPERIMENTS

A two-passes procedure is applied: 1. an acoustic-phonetic decoding is performed on the basis of which each time frame of the input utterance is given a phone label, 2. the phonetic segments are the re-scored by the four confidence measures introduced above (*GMM/LLR*, *K-nn/LLR*, *LPP* and *MP*)

### *5.1. Acoustic-phonetic Decoding*

The experiments are performed on the phonetically-labelled TIMIT database. The silence segments have been extracted (this accounts for a 3-4% fall in the recognition performances compared to the published results for the same conditions).

The data sets proportion are: -192 sentences, 6,733 phones and 48,622 frames in the core-test, -3696 sentences, 130,908 phones and 950,800 frames in the training set. Computed each centi-second, a vector-frame is represented by 12 Mel-Frequency Cepstrum Coefficients (MFCC), the energy coefficient and their first order derivates. The system comprises 46 3-states left-to-right HMMs, modelling the phonetic classes. Each HMM state encompasses a 32-GMM. The accuracy rate for this system is 63.9% on the core-test (silences excluded) for the K.-F. Lee's 39 phonetic classes paradigm [9]

To assess the performance of a test statistic as a predictor of truth or falsity, a set of tagged hypotheses is required. In this purpose, the recognizer's outputs were marked as correct or incorrect against a forced alignment of the reference t