Frame-synchronous and Local Confidence Measures for Automatic Speech Recognition

JOSEPH RAZIK\textsuperscript{1,2}, ODILE MELLA\textsuperscript{1}, DOMINIQUE FOHR\textsuperscript{1}, JEAN-PAUL HATON\textsuperscript{1}
\textsuperscript{1} LORIA-INRIA, Speech team, BP 239, 54506 Vandoeuvre-lès-Nancy, France
\texttt{givenname.familyname@loria.fr}
\textsuperscript{2} LSIS, DYNI team, Universitè du Sud Toulon-Var, BP 20132, 83957 La Garde, France
\texttt{Joseph.Razik@univ-tln.fr}

In this paper, we introduce two new confidence measures for large vocabulary speech recognition systems. The major feature of these measures is that they can be computed without waiting for the end of the audio stream. We proposed two kinds of confidence measures: frame-synchronous and local. The frame-synchronous ones can be computed as soon as a frame is processed by the recognition engine and are based on a likelihood ratio. The local measures estimate a local posterior probability in the vicinity of the word to analyse. We evaluated our confidence measures within the framework of the automatic transcription of French broadcast news with the EER criterion. Our local measures achieved results very close to the best state-of-the-art measure (EER of 23% compared to 22.0%). We then conducted a preliminary experiment to assess the contribution of our confidence measure in improving the comprehension of an automatic transcription for the hearing impaired. We introduced several modalities to highlight words of low confidence in this transcription. We showed that these modalities used with our local confidence measure improved the comprehension of automatic transcription.

Keywords: Confidence measure, likelihood ratio, posterior probability, frame-synchronous measure, speech recognition

1. Introduction

As with other pattern recognition fields, the accuracy of automatic speech recognition (ASR) systems improves each year. However, some errors (substitutions, insertions, deletions) still remain. When reading the word sequence of a recognised sentence, it is sometimes not easy to get the actual meaning of this sentence. This problem can be disastrous, for instance for deaf people watching and reading live broadcast programs subtitled or captioned by an ASR system. Two questions arise: how to detect recognition errors and how to deal with the errors found in recognised sentences? One way to solve the first point is to use confidence measures. A confidence measure estimates the level of confidence we can have in the solution.
found by the speech recognition system.

This study focuses on confidence measures that can be used for on-the-fly applications, such as automatic transcription of live programs or lessons in a classroom. In this case, it is not possible to wait for the end of the recognition process to compute the level of confidence of a word. As far as we know, the state of the art does not provide suitable confidence measures for such tasks.

We aimed to specify confidence measures that take into account the following constraints: estimated at the word level, usable for large vocabulary recognition and for endless audio stream. We have defined two kinds of measures. The framesynchronous measures can be computed as soon as the recognition engine processes a frame. The local measures work on a local view of the audio stream and need only a slight delay before being computed. Both confidence measures were evaluated in an automatic broadcast news transcription. We also conducted a preliminary study to verify the effectiveness of confidence measures for improving the comprehension of an automatic transcription for the hearing impaired.

The organisation of this paper is as follows. In the second section, we review confidence measures in speech recognition. In the third section, we introduce the confidence measures we have defined. Section 4 describes the evaluation of our confidence measures for broadcast news transcription. Finally we focus on a preliminary experiment to find out if our confidence measures could improve the comprehension of an automatic transcription by hearing impaired people.

2. Confidence measures in speech recognition

2.1. Introduction

The purpose of confidence measures is to estimate the quality of a result. In speech recognition, confidence measures are applied in various manners. For example, in keyword spotting applications confidence measures should help to decide whether to keep or reject a hypothesised keyword.\(^1\)\(^2\) Confidence measures can also be useful in detecting out-of-vocabulary words.\(^3\) Moreover, for acoustic model adaptation, confidence measure can help to select only the reliable tokens (phone, word, sentence, etc.), i.e., those with a high confidence value.\(^4\)\(^5\) In the same way, confidence measures could be used for the unsupervised training of acoustic models.\(^6\)\(^7\) They can also be used to guide the dialogue in answering services in order to require a confirmation only for words with a low confidence value.\(^8\)\(^9\) Confidence measures can be classified according to the criteria which they are based on:

- **semantic**: semantic similarity measures based on Latent Semantic Analysis (LSA),\(^10\) or Mutual information (MI),\(^11\) can be defined from the analysis of the co-occurrence of words in a set of documents.
- **language modelling**: some confidence measures are the combination of forward and backward language models;\(^12\) others use the degree of backoff in the language model estimation.\(^13\)\(^7\)
acoustic stability: this criterion lies in the idea that a word often recognised at the same position in a set of alternative hypothesis sentences should be correct. Studies differ how alternative hypotheses were chosen. However, in Ref. 17, Wessel et al. showed that this criterion was outperformed by the posterior probability based confidence measure in the framework of broadcast news transcription.

hypothesis density: this criterion represents the number of active hypotheses at a given frame through the decoding process.

duration: several studies were founded on the duration of phones in the word, or the duration of words;

likelihood ratio: confidence measures based on a likelihood ratio principle are an extension of statistical hypothesis testing which consists in accepting or rejecting an hypothesis.

lattice-based posterior probability: by its inner definition, the posterior probability seems to be a good confidence measure.

Let us consider the confidence measures summarised above according to the constraints defined in section 1. Among the confidence measures presented, acoustic stability methods need several decoding passes and so do not fit on-the-fly constraints. Confidence measures based only on semantic information or language modelling did not perform as well as those based on posterior probability. Moreover, studies merging posterior probability and purely linguistic criteria showed a small improvement compared to the posterior probability alone. Duration criteria can be computed without waiting for the end of the recognition process, but are not efficient when they are used alone.

Several works showed that confidence measures based on posterior probability outperform other kinds of measures. However, the different estimations of the posterior probability need the whole recognition process to be completed. Therefore, we aimed to define a confidence measure based on posterior probability that could be computed as soon as possible during the recognition process, and above all, without waiting for the recognition of a whole sentence.

Different studies have shown that likelihood ratio-based measures give satisfactory performance but they were evaluated for small vocabulary ASR. On the other hand, the likelihood ratio can be considered as a rough approximation of posterior probability and it can be computed frame-synchronously. Therefore, we wanted to study likelihood ratio measures in the framework of our large-vocabulary task.

Before introducing our confidence measures, we will review the likelihood ratio and posterior probability measures in the literature.

2.2. Likelihood ratio criteria

Likelihood ratio is a statistical hypothesis testing which consists in accepting or rejecting an hypothesis. This principle is widely used not only in speech recognition but also in pattern recognition. In the case of speech recognition, the hypothe-
ses are: the result of the recognition system is correct (the null hypothesis $H_0$) and the result of the system is incorrect (the alternative hypothesis $H_1$). Two kinds of error are then defined: “false” rejection of $H_0$ (Type I error) and “false” acceptance of $H_0$ (Type II error). Testing hypothesis $H_0$ against hypothesis $H_1$ is equivalent to finding out if we should accept or reject $H_0$. The Neyman-Pearson lemma says that the optimal solution of hypothesis testing is based on a likelihood ratio and a threshold $\tau$ by the following equation:

$$LR = \frac{P(X|H_0)}{P(X|H_1)}.$$  \hspace{1cm} (1)

$X$ stands for the recognition result. If $LR \geq \tau$, then $H_0$ hypothesis is accepted, otherwise it is rejected. Varying the value of $\tau$ changes the numbers of both Type I and Type II errors. It is then possible to favour one of the two kinds of error.

Confidence measures have been defined from the likelihood ratio $LR$. Let $O$ be the sequence of observations associated with a speech signal, $M$ the recognised model and $\bar{M}$ the alternative model. Hypothes $H_0$ becomes: $O$ was generated by the model $M$ and $H_1$ becomes: $O$ was generated by the alternative model $\bar{M}$. Thus equation 1 becomes: $LR = P(O|M)/P(O|\bar{M})$.

In speech recognition, $LR$ is used to estimate the reliability of a recognised unit (from frame to sentence). As we aim to estimate the confidence of recognised words, we limit this review to likelihood ratio of a word. In practice, LR can be either computed straight at the word level or as the average of ratios computed at a sub-word level (frame, phone).

Three major approaches have been proposed to model the alternative hypothesis $\bar{M}$: anti-model, catch-all model and competing models.

The most common method consists in training a specific anti-model for each model $M$. The anti-model is trained on each sample of the corpus which was not used for training the model $M$. Regarding word confidence, word anti-models are only used for very small vocabulary tasks; otherwise, the anti-models are defined at the frame level or at the phone level. Nevertheless, the confidence measures based on catch-all models are not very frequently used in ASR, even less for large vocabulary tasks.

The third approach based on competing models only utilises the models used by the recognition system. $P(O|\bar{M})$ can be rewritten as the sum of the likelihood of all the competing models:

$$LR = \frac{P(O|M)}{\sum_{\bar{M} \in V \setminus \{M\}} P(O|\bar{M})}.$$  \hspace{1cm} (2)
In the case of a large vocabulary, this method becomes virtually unfeasible because of the large number of models to take into account. One solution has been to apply competing models at the phone level.\textsuperscript{32,33,34} We have proposed another solution: to consider as competing words only the words found in the word graph generated by the recognition engine. Moreover, the word-based likelihood ratio can be computed without waiting for the end of the audio signal.

So, we have defined frame-synchronous measures based on a likelihood ratio using competing models in the framework of large vocabulary ASR.

\subsection*{2.3. Posterior probability criteria}

The posterior probability $P(W|O)$ of a sentence $W$ for the sequence of observations $O$ is defined by: $P(W|O) = P(O|W)P(W)/P(O)$. The quantity $P(O|W)P(W)$ is computed by the recognition engine during the decoding of $O$. By its inner definition, the posterior probability seems to be a good confidence measure of $W$. In the same way, we can define the confidence of a word $w$ by its posterior probability. This can be estimated by adding up the posterior probability of all the sentences which contain $w$ at the same position in the sentence.

Works dealing with posterior probability as confidence measure build the posterior probability estimation on one of the three lattice structures: a word graph, an n-best sentences list or a confusion network.

\subsubsection*{2.3.1. Word graph}

A word graph is a compact structure usually built frame-synchronously by the speech recognition engine during the decoding process. In other words, when the decoding engine processes a frame, new word hypotheses are added and linked to the current word graph. This structure holds every possible alternate word hypothesis that may end at each frame of the sentence (Figure 1). Each word hypothesis is stored with information such as: the beginning and ending times, the acoustic score, the link to the previous word which leads to the current word hypothesis considering the Viterbi algorithm, the language model score and the cumulative likelihood of the Viterbi path ending at this word.

Most of the confidence measures estimating the posterior probability from the word graph are based on the forward-backward algorithm.\textsuperscript{16,17,32,35,36}

\subsubsection*{2.3.2. N-best sentences list}

The posterior probability of $w$ can be approximated by the summation of the posterior probability of the n-best sentences which contain $w$ at the same position in the sentence.\textsuperscript{37,38,39} But the number of best sentences is limited, so this estimation is a rough approximation of the theoretical value of the posterior probability. Therefore, it is less accurate than estimations based on a word graph which contains a higher density of hypotheses.\textsuperscript{39} Furthermore, this estimation needs the recognition engine
to provide the n-best sentences, i.e., it needs the recognition process to be fully terminated. As a consequence, using n-best lists is not convenient for on-the-fly tasks.

![Word Graph Example](Fig. 1. Example of a word graph)

### 2.3.3. Confusion network

In Ref. 40, Mangu et al. have introduced a new structure, called confusion network (or sausage), more compact than a word graph in order to improve speech recognition by minimising the word error rate instead of the sentence error rate. Confusion networks are a perspicuous representation of alternative hypotheses found by the recogniser. A confusion network is elaborated from a word graph in several steps: word pruning, computing of the word posterior probability, state ordering and word clustering. Word posterior probabilities can be then reestimated from the confusion network. This process is not adapted for on-the-fly applications because it is necessary to wait until the end of the sentence. Moreover, as confusion networks require the computation of posterior probabilities from the initial word graph, it is not relevant to build confusion networks only to compute word posterior probability. Besides, in Ref. 32, Falavigna et al. compared posterior probability estimated on word graph and posterior probability estimated on confusion network. Both methods achieved a similar performance.

For these reasons, we have chosen to use word graphs to compute word posterior probability. Moreover, word graphs are built frame-synchronously during the recognition step and consequently fit our live-stream constraints.

### 3. New On-the-fly Confidence Measures

#### 3.1. Frame-synchronous measures

Our frame-synchronous confidence measures use only the data available up to the speech frame currently being processed by the recognition engine. Thus, as soon as
a frame of signal is processed by the engine, a confidence value can be computed for all the words ending at this frame.

These measures rely on the same computation pattern: a likelihood ratio between the word for which we want to evaluate the confidence (named current word henceforth in the article) and the competing words found within the word graph. The term “word” means a lexical entry. We introduce a relaxation rate $\varepsilon$ to have a more flexible selection of competing words. $\varepsilon$ corresponds to a percentage of the duration of the current word.

Let $[w, \tau, t]$ be the current word $w$ with starting time $\tau$ and ending time $t$. The word $[w', \tau', t']$ is a competing word of $w$ if:

$$\tau - \varepsilon d \leq \tau' \leq \tau + \varepsilon d.$$  
(3)

$$t - \varepsilon d \leq t' \leq t.$$  
(4)

$$(1 - \varepsilon) d \leq d' \leq (1 + \varepsilon) d.$$  
(5)

with $d = t - \tau + 1$ and $d' = t' - \tau' + 1$.  
(6)

Note that Equation (4) limits the ending time of the competing words to the ending time $t$ of the current word. This constraint is implied by the frame-synchronous characteristics of our measures: no data posterior to $t$ is available.

Let $E$ be the set of competing words of the word graph satisfying the four previous equations. Note that the current word itself belongs to the set $E$. In this way the defined likelihood ratio is normalised.

Thus, the generic pattern of the frame-synchronous measures is as follows:

$$C([w, \tau, t]) = \frac{P(O|w)P(w)}{\sum_{w' \in E} P(O|w')P(w')}.$$  
(7)

where $P(O|w)$ is the acoustic probability of the observation sequence $O$ given the word $w$, and $P(w)$ is the linguistic probability of $w$.

We have defined three different frame-synchronous measures according to the language model involved: unigram, bigram or trigram.

3.1.1. Multiple occurrence management

Introducing a relaxation rate to select competing words implies managing multiple occurrences of the same word with close beginning and ending times. We have chosen two methods to deal with this issue:

- a summation method adding up the likelihood of every occurrence $[w, \tilde{\tau}, \tilde{t}]$ of the current word $w$, and adding up the likelihood of every occurrence $[w', \tau', t']$ of the competing words $w'$

$$C([w, \tau, t]) = \frac{\sum_{{[w, \tilde{\tau}, \tilde{t}] \in E}} P(O|w)P(w)}{\sum_{w'} \sum_{[w', \tilde{\tau}', \tilde{t}'] \in E} P(O|w')P(w')}.$$  
(8)
a maximisation method keeping only the occurrence with the maximal acoustic score. Let $[w, \hat{\tau}, \hat{t}]$ be this occurrence. Let $\hat{E}$ be the sub-set of $E$ containing one occurrence per competing word $w'$, the occurrence $[w', \hat{\tau}', \hat{t}]$ which maximises acoustic score.

$$C([w, \tau, t]) = \frac{P(o_\tau^t | w)P(w)}{\sum_{[w',\hat{\tau}',\hat{t}'] \in \hat{E}} P(o_{\hat{\tau}'}^{\hat{t}'} | w')P(w')}.$$  

(9)

For each frame-synchronous measure, experiments were conducted for the summation method and the maximisation method but for the sake of simplicity only the equations for the maximisation method will be given in the following sections.

### 3.1.2. Unigram measure

This confidence measure uses only basic and highly local information: acoustic scores and unigram language probabilities. As acoustic scores vary in a very large dynamic range of values, two scaling factors $\alpha$ and $\beta$ are introduced for respectively acoustic and linguistic probabilities.

With the maximisation method, the confidence measure of $w$ is then defined by:

$$C([w, \tau, t]) = \frac{p(o_\tau^t | w)^\alpha p(w)^\beta}{\sum_{[w',\hat{\tau}',\hat{t}'] \in \hat{E}} p(o_{\hat{\tau}'}^{\hat{t}'} | w')p(w')^\beta}.$$  

(10)

### 3.1.3. Bigram measure

The previous confidence measure is modified in order to take into account the word neighbourhood by using the bigram language probabilities with the previous words.

With the maximisation method, the bigram probabilities are computed between the occurrence $[w, \hat{\tau}, \hat{t}]$ and its previous words $w_p$. Respectively, the bigram probabilities of the words $w'$ are computed with the previous words $w'_p$ of the occurrences $[w', \hat{\tau}', \hat{t}']$ of $\hat{E}$. The confidence measure is then defined by:

$$C([w, \tau, t]) = \frac{p(o_\tau^t | w)^\alpha (\sum_{w_p} p(w | w_p)p(w_p))^\beta}{\sum_{[w',\hat{\tau}',\hat{t}'] \in \hat{E}} p(o_{\hat{\tau}'}^{\hat{t}'} | w')^{\alpha}(\sum_{w'_p} p(w' | w'_p)p(w'_p))^{\beta}}.$$  

(11)

### 3.1.4. Trigram measure

The vicinity in the past can be further enlarged through the introduction of a trigram language model. The confidence measure is still frame-synchronous because the trigram probability is computed with words previous to the current word.

In the case of the maximisation method, for each maximal occurrence $[w, \hat{\tau}, \hat{t}]$, we consider its previous words $[w_p, \tau_p, t_p]$ and the previous words $[w_{pp}, \tau_{pp}, t_{pp}]$ of
that contain \([w_p, \tau_p, t_p]\) according to the two ways defined for the bigram measure. By applying the same considerations to the words \(w', w'_p\) and \(w''_p\), the trigram measure is defined as:

\[
C([w, \tau, t]) = \sum_{[w', \tau', t'] \in E} p(o_{\hat{w} | w'})^\alpha (\sum_{w_p, w_{pp}} p(w | w_p, w_{pp}) p(w_p | w_{pp}) p(w_{pp}))^{\beta}.
\]

3.2. Local measures

Our local measures are based on an estimation of the posterior probability of a word. They can use data slightly posterior to the current word. However, this data is limited to the local vicinity of this word and the confidence estimation does not need the recognition of the whole sentence. Thus, only a short delay is needed to allow the data to be available for computing the measures.

For the current word, let us define a neighbourhood \(V\). The scope of \(V\) is given by a fixed amount of frames before (past neighbourhood of \(x\) frames) and after the word (future neighbourhood of \(y\) frames). The delay before the computation of the local measure depends on the duration of the future neighbourhood. From \(V\), we extract the corresponding sub-graph \(SG\) from the word graph provided by the recognition engine. For a given neighbourhood which begins at frame \(b\) and ends at frame \(e\), a word \([w, \tau, t]\) is included in the sub-graph if \(b \leq \tau\) and \(t \leq e\).

We then compute the local posterior probability of \(w\) on \(SG\) by using the following forward-backward method.

Let \(p([w, \tau, t] | o_b^\alpha)\) be the posterior probability of a sequence of \(M\) word hypotheses \(w_1, \ldots, w_M\), given the acoustic observations \(o_b, \ldots, o_e\).

We define the local posterior probability of a word hypothesis \([w, \tau, t]\), noted \(p([w, \tau, t] | o_b^\alpha)\), as the sum of the posterior probability of all the word sequences of \(SG\) that contain \([w, \tau, t]\).

By analogy with the forward-backward algorithm used at a word level, for a word hypothesis \([w, \tau, t]\) we computed the forward probability \(\Phi([w, \tau, t])\) and the backward probability \(\Psi([w, \tau, t])\) but using only words belonging to \(SG\).

The forward and backward probabilities of words can be computed recursively, Equation (13) represents the recursive forward probability formula, and Equation (14) the recursive backward probability using a bigram language model.

\[
\Phi([w, \tau, t]) = p(o_{\hat{w} | w})^\alpha \sum_{w_p, \tau'} \Phi([w_p, \tau', \tau - 1]) p(w | w_p)^\beta.
\]

\[
\Psi([w, \tau, t]) = p(o_{\hat{w} | w})^\alpha \sum_{w_n, \tau'} \Psi([w_n, t + 1, \tau']) p(w_n | w)^\beta.
\]

In Equation (13), \([w_p, \tau', \tau - 1]\) denotes any previous word in \(SG\) that ends at \(\tau - 1\). In Equation (14) \([w_n, t + 1, \tau']\) denotes any following word in \(SG\) that begins at \(t + 1\). \(\alpha\) and \(\beta\) are scaling factors for the acoustic and the language model score.
The local posterior probability of the word \([w, \tau, t]\), using forward and backward probabilities, is given by:

\[
p([w, \tau, t]|o_b) = \Phi([w, \tau, t])\Psi([w, \tau, t])p(o_b)p(o_t|w)^{\alpha}.
\]  

(15)

One major point in the computation of the posterior probability is the estimation of \(p(o_b)\). This quantity can be obtained with the forward probability as:

\[
p(o_b) = \sum_w \sum_{\tau} \Phi([w, \tau, e]).
\]  

(16)

In the extracted sub-graph associated with \(V\), several occurrences of the current word may occur almost simultaneously. The previous forward-backward method computes the posterior probability of each of these occurrences. Keeping only one occurrence as confidence measure underestimates the true posterior probability of the word. Thus, a flexibility factor \(\eta\) was introduced and the posterior probability of each occurrence of the current word satisfying several criteria according to \(\eta\) was added. More precisely, we defined a set \(F\) of occurrences \([w, \tilde{\tau}, \tilde{t}]\) of the word \([w, \tau, t]\) belonging to the sub-graph and satisfying the following constraints:

\[
\tau - \eta d \leq \tilde{\tau} \leq \tau + \eta d. 
\]  

(17)

\[
t - \eta d \leq \tilde{t} \leq t + \eta d. 
\]  

(18)

\[
(1 - \eta) d \leq \tilde{d} \leq (1 + \eta) d. 
\]  

(19)

\(\eta\) represents a proportion of the duration \(d = t - \tau + 1\) of the word \(w\).

The confidence measure of the current word \(w\) is then defined as the sum of the local posterior probability of each occurrence of \(w\) in \(F\):

\[
C([w, \tau, t]) = \sum_{[w, \tilde{\tau}, \tilde{t}] \in F} p([w, \tilde{\tau}, \tilde{t}]|o_b^\eta).
\]  

(20)

**Note:** Equations (17) to (19) seem to be similar to equations (3) to (5), but the flexibility factor \(\eta\) plays a totally different role than the relaxation rate \(\varepsilon\). In the frame-synchronous measures case, \(\varepsilon\) is used to define a set of competing hypotheses in order to compute a likelihood ratio. In the local measures case, \(\eta\) makes it possible to take into account the multiple occurrences of the current word within the neighbourhood \(V\), then to add up their posterior probabilities.

We have defined frame-synchronous and local confidence measures. In the next sections, we will assess them in the framework of two tasks.

**4. Evaluation of our Confidence Measures for Broadcast News Transcription**

We first evaluated our confidence measures in the framework of automatic broadcast news transcription. We optimised our confidence measures on a development corpus and determined for each one an optimal decision threshold corresponding to the best
Equal Error Rate (ERR) criterion. We compared their performance to a state-of-the-art measure (reference measure). Finally, we assessed our measures on a test corpus and computed the false acceptance (FA) and false rejection (FR) rates.

4.1. Equal Error Rate

To evaluate confidence measures, it is necessary to consider if a recognised word is correct or incorrect. In our experiments, a recognised word is correct if it matches the reference word exactly, i.e., we considered every mistake, orthographic or grammatical, as an error. For instance, given the uttered words “three cars”, the two following recognised phrases are considered as wrong: “thr ee car” and “three bars”. “three car” is grammatically wrong and, even if the reader will correct it, as our application is intended for students, grammatical mistakes are not allowed.

A word is labelled accepted if its confidence value is greater than a decision threshold; otherwise, the word is labelled as rejected.

The EER evaluation method requires the computation of the false acceptance (FA), false rejection (FR), defined as follows:

\[
FA = \frac{\text{number of incorrect words labelled as accepted}}{\text{number of incorrect words}}.
\]

\[
FR = \frac{\text{number of correct words labelled as rejected}}{\text{number of correct words}}.
\]

These rates vary according to the value of the decision threshold. The EER corresponds to the threshold value for which the false acceptance and false rejection rates are equal. The decision threshold was determined on the development corpus.

4.2. Experimentation conditions

4.2.1. Automatic News Transcription System ANTS

For our study, we used our large vocabulary speech recognition system ANTS. This system is based on Julius, developed by researchers at Kyoto University. Our measures were computed on the internal word graph generated by the frame-synchronous first pass of Julius. The word graph contains an average of 470 word hypotheses per frame on the development corpus.

4.2.2. Acoustics and language models

Most of the speech recognition systems are based on HMM models even if new stochastic models begin to be applied to speech recognition (Bayesian Networks, Conditional Random Fields). Acoustic HMM models of ANTS were trained on a corpus of about 40 hours of French radio broadcast news. This corpus was extracted from a larger corpus provided by the 2006 ESTER French evaluation campaign. This training corpus contains only broadband speech (no narrow band, no music
segments). The parametrisation was based on MFCC (Mel Frequency Cepstral Coefficient),\textsuperscript{46} with CMS (Cepstral Mean Subtraction) normalisation.

Bigram and trigram language models were trained with the CMU Toolkit\textsuperscript{47} on 16 years of the “Le Monde” French newspaper, completed by the manual transcription of the broadcast news training corpus. A total of 2.5M bigrams and 5.8M trigrams was estimated. The lexicon contains 54747 words.

4.2.3. Development and test corpora

53 minutes of French broadcast news programs was used as a development corpus for the decision threshold and for tuning the relaxation/flexibility rates and the scaling factors. The test corpus is composed of 56 minutes of broadcast news programs. Both corpora contain about 11000 words and 11.5 words per sentence on average. This represents an average of 330 frames per sentence according to a 10 ms frame duration. The word error rate was 31% on the development corpus and 33% on the test corpus. Neither speaker diarization nor speaker adaptation were performed due to live constraints. Both corpora contained broadband speech, no telephone speech and no music. However, several sentences had background noise or music.

According to the corpus size, the different results are given to a significance confidence of around 0.8% at the 0.95 level of significance.

4.3. Reference measure

In this work, we assessed and compared our confidence measures with each other as well as with a reference measure. We chose as reference measure a state-of-the-art measure based on the estimation of the posterior probability computed on the whole signal, once the speech recognition engine had processed the entire sentence. The reference measure of the current current word $w$ with starting time $\tau$ and ending time $t$ is the sum of the posterior probability of each occurrence $w$ in $G$:

$$C([w, \tau, t]) = \sum_{[w, \hat{\tau}, \hat{t}] \in G} p([w, \hat{\tau}, \hat{t}]|o^T).$$  \hspace{1cm} (23)

$G$ is the set of occurrences of $w$ belonging to the whole word graph and satisfying equations (17) to (19). The posterior probability of each occurrence of $w$ is computed using the forward-backward method.

This reference measure is currently known to be one of the most accurate.\textsuperscript{17} As our confidence measures had only a partial knowledge of the audio signal, the performance obtained by this reference measure could be considered as a limit for our measures. Therefore, we first evaluated the reference confidence measure on the development corpus which obtained an EER of 22.0%.

4.4. Frame-synchronous measures

Preliminary experiments showed that for all frame-synchronous measures, the maximisation method provided better or equal results than the summation method but
the difference was not significant (cf. Table 4). Thus, in following experiments, only the results for maximisation method will be presented. We then optimised scaling factors \((\alpha, \beta)\) and relaxation rate \(\epsilon\) on the development corpus for the unigram, bigram and trigram measures, described in section 3.1.

Table 1 shows the results for the bigram measure using the maximisation method to manage the multiple occurrences. The best results were obtained with a small relaxation rate (10%). Let us remember that the higher the relaxation rate is, the more competing words are taken into account for the computation of the likelihood ratio. As expected, if the relaxation rate is too high, the measure compares the likelihood of the current word with the likelihood of words that are not really competing words for the recognition engine. These words were too time shifted. On the other hand, we can observe that the \(\beta/\alpha\) ratio had little effect on the EER.

Table 1. EER evolution with different scaling factors ratios and relaxation rates for the bigram frame-synchronous measure

<table>
<thead>
<tr>
<th>(\epsilon)</th>
<th>(\beta/\alpha) ratio</th>
<th>1</th>
<th>5</th>
<th>9.5</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>39.7%</td>
<td>38.8%</td>
<td>38.4%</td>
<td>38.1%</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>38.9%</td>
<td>37.5%</td>
<td><strong>37.4%</strong></td>
<td>37.8%</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>40.6%</td>
<td>38.8%</td>
<td>38.6%</td>
<td>38.9%</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>42.4%</td>
<td>40.5%</td>
<td>40.3%</td>
<td>39.6%</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>45.3%</td>
<td>42.7%</td>
<td>42.0%</td>
<td>41.3%</td>
<td></td>
</tr>
</tbody>
</table>

On the development corpus, the optimal relaxation rate was the same for each of our frame-synchronous measures: \(\epsilon = 0.1\).

Table 2 summarises the evolution of the EER with regard to the \(\beta/\alpha\) ratio for our three major frame-synchronous confidence measures: unigram, bigram and trigram. From these results, we can notice that using an n-gram language model rather than an (n-1)-gram language model introduced a slight improvement, but not significant. As expected, the performance of these measures is worse than the reference measure because they are very local, and moreover the likelihood ratio is a drastic approximation of the posterior probability. But, these confidence measures have the advantage of being simple and fully frame-synchronous.

Table 2. EER evolution with different scaling factors ratios for the unigram, bigram and frame-synchronous measure \((\epsilon = 0.1)\)

<table>
<thead>
<tr>
<th>confidence measure</th>
<th>(\beta/\alpha) ratio</th>
<th>1</th>
<th>5</th>
<th>9.5</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigram</td>
<td>39.8%</td>
<td><strong>37.6%</strong></td>
<td>38.4%</td>
<td>38.2%</td>
<td></td>
</tr>
<tr>
<td>bigram</td>
<td>38.9%</td>
<td>37.5%</td>
<td><strong>37.4%</strong></td>
<td>37.8%</td>
<td></td>
</tr>
<tr>
<td>trigram</td>
<td>38.7%</td>
<td>37.2%</td>
<td><strong>37.1%</strong></td>
<td>38.3%</td>
<td></td>
</tr>
</tbody>
</table>
4.5. **Local measures**

In this section, we evaluate our local posterior probability measures (section 3.2) using the same EER criterion on the development corpus.

4.5.1. **Symmetric neighbourhood**

The symmetric neighbourhood is defined by the number of frames added before and after the **current word**. For example if we consider a word \(w\) of 30 frames duration and a **84-frame symmetric neighbourhood**, the total number of frames of the partial sub-graph \(SG\) is equal to 198 frames (with a 10ms frameshift).

The scaling factors \(\alpha\) and \(\beta\), and the flexibility factor \(\eta\) were tuned on the development corpus for a fixed neighbourhood size of 84 frames. The EER results are shown in Table 3. The best EER is obtained with \(\alpha = 0.1, \beta = 0.95, \eta = 0.5\).

Table 3. EER evolution according to different scaling ratios and flexibility factors for the local symmetric posterior probability measure with neighbourhood size = 84 frames.

<table>
<thead>
<tr>
<th>(\eta)</th>
<th>(\beta/\alpha) ratio</th>
<th>1</th>
<th>9.5</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>36.3%</td>
<td>28.5%</td>
<td>28.8%</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>35.8%</td>
<td>23.8%</td>
<td>27.1%</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>35.7%</td>
<td><strong>23.0%</strong></td>
<td>26.9%</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>35.7%</td>
<td><strong>23.0%</strong></td>
<td>26.7%</td>
<td></td>
</tr>
</tbody>
</table>

In order to study the influence of the size of the neighbourhood, we plotted the evolution of the EER versus this size with \(\alpha = 0.1, \beta = 0.95, \eta = 0.5\) on the development corpus. Figure 2 shows that up to 84 frames the EER sharply decreases to a value close to the EER of the reference measure. Above 84 frames, the influence of the neighbourhood size is weaker. Compared to the reference posterior probability measure calculated on the whole sentence, the local measure with an 84-frame neighbourhood provides a close score: 23.0% for the local measure versus 22.0% for the reference measure. We can note that 84 frames corresponds to the average duration of two consecutive subsequent words in the development corpus.

![Fig. 2. EER according to the duration of the symmetric neighbourhood \((\alpha, \beta, \eta) = (0.1, 0.95, 0.5)\)](image-url)
4.5.2. Asymmetric neighbourhood

In this section, we investigate the EER behaviour of our local confidence measure with an asymmetric neighbourhood. The duration of the past and future neighbourhoods are independently defined. This allows more data from the past (already processed, so available) to be used without increasing the delay introduced by the future neighbourhood.

Figure 3 shows the EER evolution versus the duration of the past neighbourhood which runs from 40 to 200 frames. Three curves are plotted according to the duration of the future neighbourhood: 40, 60 and 84 frames. We can observe that the shape of the curves is quite the same for the 3 durations.

As expected, we can note that the larger the neighbourhoods are, the better the local measures perform. In particular, for a given duration of the future neighbourhood, increasing the duration of the past neighbourhood up to 84 frames dramatically improved the EER. After 84 frames, the improvement was slighter.

For a future neighbourhood of only 60 frames we obtained a performance close to the reference measure (23.5%) by increasing the past neighbourhood to 200 frames. In this case, the time delay is of only about half a second.

If the future neighbourhood is decreased to 0 frames, this measure becomes frame-synchronous and is still based on a local estimation of the posterior probability. If we take into account a 200-frame past neighbourhood, this frame-synchronous measure achieves an EER of 30.5%. This value is significantly better than the best EER obtained by our frame-synchronous measures based on a likelihood ratio (37.0%). This can be observed in DET (Detection Error Tradeoff) curves which plot FA and FR rates by varying the decision threshold in Figure 4.

Several explanations can be given:
Fig. 4. DET curves for reference, frame-synchronous trigram, and two local confidence measures

- more data is taken into account by posterior probability based measure,
- the likelihood ratio criterion is less accurate because only the acoustic likelihood of the current word is taken into account,
- the bigram measure (for instance) calculates a ratio between sequences of exactly two words, regardless of their duration. The local estimation of the posterior probability by the forward-backward algorithm does not compare sequences with the same number of words but with the same duration.

4.5.3. Confidence value vs. correct word rate

It would be interesting to study the correlation between the computed confidence values and the results provided by the recognition system. The mean confidence value should be close to the correct word rate (CWR), otherwise the confidence measure will over- or underestimate the reliability of the words.

To analyse this correlation, we defined a method based on the following step:
- computing the confidence value of the recognised words,
- sorting the words according to their confidence value,
- splitting the sorted words in $N$ sets of the same size,
- for each set, computing the mean confidence values and the CWR.

For $N = 20$ sets, Figure 5 shows the evolution of the confidence value and of the CWR on the development corpus for the local 84-84 measure. The plain curve shows the mean confidence values and the dashed curve plots the CWR computed for each set. As both curves increase, we can conclude that our local confidence measure succeeded in capturing confidence information. However, the figure also shows that it globally overestimated the confidence of words for which the confidence value is less than 0.5.
4.6. Results on the test corpus and discussion

After optimising our confidence measure on the development corpus, we assessed them on the test corpus. Table 4 summarises the results according to EER on the development corpus and according to FA, FR on the test corpus. In order to have one value for the experiment on the test corpus, we computed the average of FA and FR, even if the decision threshold was not tuned for this rate.

Table 4. Results of the reference measure and our confidence measures (local and frame-synchronous) on the development corpus (EER) and on the test corpus (FA, FR)

<table>
<thead>
<tr>
<th>measure</th>
<th>Dev. corpus</th>
<th>Test corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>FR</td>
</tr>
<tr>
<td>reference</td>
<td>22.0%</td>
<td>21.2%</td>
</tr>
<tr>
<td>local 84-84</td>
<td>23.0%</td>
<td>23.7%</td>
</tr>
<tr>
<td>local 200-60</td>
<td>23.5%</td>
<td>24.9%</td>
</tr>
<tr>
<td>local 60-60</td>
<td>25.5%</td>
<td>27.3%</td>
</tr>
<tr>
<td>local 200-0</td>
<td>30.5%</td>
<td>29.1%</td>
</tr>
<tr>
<td>trigram with maximisation</td>
<td>37.1%</td>
<td>34.5%</td>
</tr>
<tr>
<td>bigram with summation</td>
<td>37.4%</td>
<td>36.2%</td>
</tr>
<tr>
<td>bigram with maximisation</td>
<td>37.4%</td>
<td>36.6%</td>
</tr>
<tr>
<td>unigram</td>
<td>37.6%</td>
<td>38.8%</td>
</tr>
</tbody>
</table>

We can note that the results obtained by our confidence measures have the same behaviour on the development corpus and on the test corpus. The measures we defined are stable on both corpora.
For frame-synchronous measures, we can note that a higher order of n-gram involved in the definition of the measure provides better results. The frame-synchronous measures obtained worse results than local confidence measures on the test corpus as well as on the development corpus. This could be expected because the likelihood ratio is too rough an approximation of the posterior probability.

On both corpora, the local confidence measures based on a local estimation of the posterior probability provided performance close to the reference measure which is also an estimation of the posterior probability, but computed on the whole sentence.

The local measure with a symmetric neighbourhood of 84 frames (0.84s) achieved an EER rate close to the reference measure on both corpora. This confidence measure using a short neighbourhood may thus be convenient for a task such as on-demand confidence checking when only part of the signal is available.

It is remarkable that a local estimation of the posterior probability of a word achieves quite the same performance as the posterior probability estimated on the whole sentence. That can be explained by the fact that a recognised word located at the beginning of a sentence has a weak influence on the recognition of a word located in the middle of the sentence.

The local frame-synchronous (200−0) measure achieved rates between those of the local measures and those of the frame-synchronous measures based on likelihood ratio.

5. Preliminary Experiment: Towards Automatic Transcription for the Hearing Impaired

In Europe, there are several million deaf or hearing impaired people with a growing number of people becoming deaf (old age, work in highly noisy conditions, excessive music amplification, etc.). The usual ways used for the deaf to communicate are Sign Language or Cued Speech. Cued speech consists of lipreading completed by a small number of handshapes (to discriminate consonants) in different locations near the mouth (to discriminate vowels). But using these languages is not satisfactory. For people becoming deaf, learning any of these additional new languages is very difficult and de-motivating. For hearing impaired students, it is difficult to follow a course in a standard school classroom. For lip reading, the students are confronted with several difficulties: distance to the teacher, teacher facing back the students to write on the blackboard. For sign language and cued speech, an additional skilled person is required in the classroom.

Automatic recognition could provide great help for the hearing impaired. For example the transcription of the teacher’s speech could be displayed on a laptop synchronously or with a slight delay with the actual talk. But, automatic speech recognition systems (ASR) are not perfect and recognition errors may generate

*http://www.cuedspeech.org
transcriptions which are very difficult for the hearing impaired to understand.

We wanted to know if highlighting the words which are perhaps wrong can improve the comprehension of the automatic transcription. We therefore conducted a preliminary experiment in order to evaluate how our confidence measures can improve the understanding of an automatic transcription, and, to assess an experimental protocol. For these reasons, no hearing impaired person was included in the preliminary experiment. The subjects were 20 students with normal hearing ability, but of course, they could not listen to the audio signal, and, they were not phonetically skilled. Indeed, involving deaf or hearing impaired people was difficult to manage for a first experimentation due to the need for a sufficient number of participants as well as the availability of the subjects and their speech therapists.

5.1. Description of the visual modalities

We hypothesised that highlighting words of low confidence level would help the reader to correct the errors made by the recognition system and to better understand the meaning of the sentence. We thus conducted an experiment by introducing new visual modalities into the automatic transcription provided by ANTS.

Automatic transcriptions of the audio signal were presented to subjects according to the following four modalities:

- **Raw**: the automatic transcription directly provided by the speech recognition system with no indication (all words in black colour);
- **Oracle**: the automatic transcription in which misrecognised words were written in blue. A word is considered as misrecognised if it differs from the reference transcription (cf. section 4.1);
- **Confidence**: the automatic transcription in which words tagged as incorrect by our confidence measure were written in blue;
- **Phonetic**: the automatic transcription in which words tagged as incorrect by our confidence measure were written in blue using a simplified phonetic alphabet. Moreover, the successive incorrect words were concatenated before being transcribed.

The Oracle modality makes it possible to assess the usefulness of a perfect confidence measure compared to a Raw transcription.

We introduced a Phonetic transcription of the words of low confidence because we think that, in this way, the reader can guess the right word by focusing on how the wrong word sounds rather than focusing on what the wrong word means. For that purpose, we used a simplified phonetic alphabet because a lot of people are not aware of the International Phonetic Alphabet. For instance, we replaced /y/ by /u/, because the grapheme ”u” is pronounced /y/ in French.

The successive incorrect words were concatenated before being transcribed in order to remove the potential wrong word segmentations found by the recognition system. Table 5 shows the four modalities for a French sentence.
5.2. Experimentation protocol

The 20 subjects were given the printed automatic transcription for which the words with low confidence level were highlighted according to four modalities and we asked the subjects to do the following tasks in 15 minutes:

- restore (guess) the original text of a part of the transcription (60 words on average);
- answer 3 or 4 questions about the meaning of another part of the transcription;
- answer a set of subjective questions about their feeling concerning the different modalities.

The first two tasks allow us to test two experimental protocols. Restoring the original text is not tractable during live transcription but it can be useful in the framework of a school classroom: students can reread the transcription after the course. This task could be considered as a baseline test. Indeed, if the confidence measure does not improve comprehension for this task, there is little chance that it will improve comprehension in the case of live scrolling transcription. On the other hand, asking questions about a live scrolling transcription is an obvious method to assess if highlighting low confidence words improves comprehension. Unfortunately, the students’ answers were difficult to interpret because the expected answers were either obvious or impossible to guess. So, we will not discuss it any further.

5.3. The chosen texts

The texts dealt with various topics: a fairy tale, a chronicle about a car race (“Le Mans”), a story about a flight expedition and an investigation story about the theft of a computer. We chose these texts because they were initially calibrated as reading tests for students entering secondary school (about 11 years old).

All the texts were read and recorded by the same speaker. The recordings were transcribed with the real-time version of the system ANTS. No adaptation was applied, either for the speaker or for the acoustic environment. This explains why the average recognition rate for the four texts was 71.4%.

For each transcribed text, we computed the confidence of each recognised word. We chose the local confidence measure with a symmetric neighbourhood of 84...
frames that achieved good performance (cf. section 4.6). We then compared the confidence value of a word with the decision threshold tuned on the development corpus (section 4.2.3) according to EER. When the confidence value was less than this decision threshold, the word was highlighted.

For each text we selected a sub-part of the automatic transcription (240 words on average) which was then modified according to the four modalities. These modified transcriptions were submitted to 20 subjects. Each subject was given each of the four texts, but with a different modality per text.

5.4. Results

5.4.1. The restored text

Table 6 shows word error rate (WER) for each restored text and each modality. The ASR column gives the WER of the automatic recognition system. We did not take spelling errors into account (i.e. words that sound and mean in the same way but that are spelled differently) because we wanted to assess the comprehension of the transcription and not if the student is a good speller. Indeed, there are a lot of cases of spelling errors in French: singular/plural (e.g. voiture vs. voitures), masculine/feminine (perdu vs. perdue), conjugation (aimait vs. aimais).

We can note that even without any additional information (Raw), the participants were able to correct some errors of the ASR. Afterwards, we compared each result with the Raw word error rate. On average, the original text was better restored if a visual clue in the word confidence was provided, either by highlighting the words that really were false (Oracle modality) or by using a confidence measure (Confidence or Phonetic modality). This showed that the introduction of confidence measures helped the reader to correct an erroneous transcription.

Table 6. WER on the texts restored by the subjects according to the different modalities

<table>
<thead>
<tr>
<th>Text</th>
<th>ASR [%]</th>
<th>Raw [%]</th>
<th>Oracle [%]</th>
<th>Confidence [%]</th>
<th>Phonetic [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le Mans</td>
<td>18.8</td>
<td>10.4</td>
<td><strong>7.8</strong></td>
<td>11.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Fairy tale</td>
<td>36.1</td>
<td>14.2</td>
<td>20.0</td>
<td>15.7</td>
<td>15.4</td>
</tr>
<tr>
<td>Flight expedition</td>
<td>47.4</td>
<td>47.4</td>
<td>40.7</td>
<td>36.5</td>
<td><strong>34.4</strong></td>
</tr>
<tr>
<td>PC’s theft</td>
<td>20.5</td>
<td>17.9</td>
<td><strong>14.4</strong></td>
<td>16.9</td>
<td>16.4</td>
</tr>
<tr>
<td>Average</td>
<td>30.7</td>
<td>22.5</td>
<td>20.7</td>
<td>20.0</td>
<td><strong>18.8</strong></td>
</tr>
</tbody>
</table>

On average, the Phonetic modality gave the best results, but the difference is not significant. There may be two reasons to explain this. Phonetically writing a low confidence word rather than just colouring it, guides the subject less toward guessing a word with the same root or with the same meaning. This hypothesis correlates with the subjects’ answers about the modalities. Another reason may be related to the fact that we concatenated successive low confidence words in the phonetic transcription case. Indeed, this concatenation could remove wrong lexical segmentation. The reader can then find the original words more easily, as in the
example shown in Table 5. The experiment was carried out in French, However this modality may not be useful for students who are too young or for people who are born deaf.

5.4.2. Subjective questions
In their answers to the subjective questions, all the subjects expressed their preference for the Phonetic modality because it helped them more in correcting the transcriptions. We should point out that the subjects were not phonetically skilled. However, the fact that the Phonetic modality was chosen as the best modality by the subjects should be validated with people who were born deaf and who may have a different relationship to some kind of phonetic memory.

Contrary to the results shown in Table 6, a majority of the subjects did not feel that highlighting misrecognised words in blue helped them. In fact they felt disturbed. For Confidence and Oracle modalities, several subjects pointed out that writing the low-confidence word in colour guided them towards a word having a similar root or meaning instead of a word that sounded the same. Indeed, in the phonetic modality the highlighted words did not disturb them.

5.5. Conclusion
Several conclusions and guidelines for future experiments can be drawn from this preliminary experiment.

First, we showed that adding confidence information, specially our local confidence measure, is useful in increasing the comprehension of recognised sentences.

We then evaluated two modalities to highlight low confidence words: either by just colouring them or by using a simplified phonetic alphabet. All the subjects preferred the Phonetic modality. Above all, the phonetic writing of potentially wrong words provides better results than just colouring them. This modality could be useful for people who had good hearing in the past and who now are deaf. However, the efficiency of this modality should be confirmed for people born deaf. Moreover, it must be confirmed by a “live” experiment with scrolling text. But, a new problem arises: how to evaluate the readers’ understanding? The restoring task is impracticable and finding relevant questions is difficult. The best solution might be to display only a few lines of the automatic transcription and to ask subjects to write what they have understood.

6. Conclusion
We designed two kinds of confidence measures that can be computed as soon as a frame is processed by the recognition engine or after a short delay. Thus, these measures can be used for on-the-fly applications and without waiting for the end of the recognition process. Frame-synchronous measures can be computed as soon as a frame is processed by the speech recognition engine and are based on a likelihood
ratio between the analysed word and competing words. We defined and evaluated several methods to select competing words. Local measures take into account information in a limited neighbourhood of the word to analyse. These measures are based on the local estimation of the posterior probability. We tested different durations for symmetric and asymmetric neighbourhoods. The future neighbourhood introduces a short delay between the ending time of the word analysed and the frame currently being processed by the engine.

We first assessed our confidence measures according to EER in the framework of automatic broadcast news transcription. Our local measures achieved a performance very close to the most accurate state-of-the-art measure based on posterior probability but computed on the whole signal (EER of 23.0% compared to 22.0%).

As expected, the measures based on the posterior probability were more accurate than those based on a likelihood ratio. Moreover, when we set the future neighbourhood to a null size, the defined local measure became frame-synchronous. With a past neighbourhood of 200 frames, this measure obtained an EER of 30.5%. This is an intermediate value between the rates of our best local measure (23.0%) and best frame-synchronous measure (37.1%). Results obtained on the test corpus confirm those obtained on the development corpus.

We then conducted a preliminary experiment to test the contribution of our confidence measure to improving the comprehension of an automatic transcription for the hearing impaired. To assess this, we chose to highlight words of low confidence either by colouring them or by phonetically writing them. Both modalities used with our local confidence measure improved the comprehension of automatic transcription. Moreover, phonetic writing of potentially wrong words provides better results than just colouring them and this modality was preferred by all subjects.

To conclude, we can note that this first experiment is interesting and encouraging because it shows that tools based on speech recognition and confidence measures can help disabled people. But before conducting an experiment with deaf students in a classroom, other experiments will be required to determine the best modality to use and to validate the assessment method.

References
24 J. Razik, O. Mella, D. Fohr and J.-P. Haton