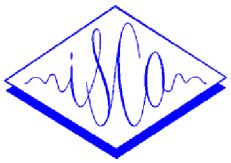

WOCCI 2008: The 1st Workshop on
Child, Computer, and Interaction
Chania, Crete, Greece
October 23, 2008

Comprehension Improvement using Local Confidence Measure: Towards Automatic Transcription for Classroom

Joseph Razik
LORIA-INRIA
Speech Group
Nancy, France
razik@loria.fr

Odile Mella
LORIA-INRIA
Speech Group
Nancy, France
mella@loria.fr

Dominique Fohr
LORIA-INRIA
Speech Group
Nancy, France
fohr@loria.fr

Jean-Paul Haton
LORIA-INRIA
Speech Group
Nancy, France
jph@loria.fr

ABSTRACT

We conducted a preliminary study to evaluate the contribution of confidence measure for improving the comprehension of an automatic transcription. The future framework is deaf children studying in a standard school classroom. We defined local confidence measures that can be estimated as soon as possible without waiting for the recognition process to be completed. We the defined different modalities to highlight words of low confidence in an automatic transcription and we presented the different modified transcriptions to several subjects who had to answer to some questions of comprehension and to restore the sentences originally uttered. We showed that highlighting words of low confidence can improve the comprehension of the automatic transcription.

1. INTRODUCTION

For deaf children who want to study in a standard school classroom, the current ways to take lessons are lip reading, sign language or cued speech. Cued speech consists in lip reading completed by a small number of handshapes (to discriminate consonants) in different locations near the mouth (to discriminate vowels)[1].
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 hard of hearing. 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. A second example could be using automatic transcription to pilot a talking head with cued speech [7]. But, automatic speech recognition systems (ASR) are not perfect and recognition errors may generate transcriptions which are very difficult to understand by the hard of hearing.

We wanted to know if highlighting the words which are perhaps wrong can improve the comprehension of the automatic transcription. Confidence measure is a well-known method to estimate if a recognised word is the right one.

Thus in this study, we first focused on defining confidence measures that could be used for on-the-fly applications, and then, on assessing how the introduction of confidence information into automatic transcription could improve understanding. For that, we conducted a preliminary experiment with 20 hearing students but who were not able to listen to the audio signal. Indeed, for a preliminary experiment, involving deaf children was too difficult.

2. CONFIDENCE MEASURE

In speech recognition, the aim of a confidence measure is to estimate the probability that the word recognised is the right one.
We defined a local confidence measure which can be used in the framework of applications working on streamed or live programs. The main idea of our local measure is to estimate the posterior probability of the analysed word on a neighbourhood around the word. A short delay is then required to allow the data to be available for computing the measure. The confidence estimation does not need to wait the end of the recognition of the whole sentence.

2.1 Definition of our Local Measure

Let $[w, \tau, t]$ stand for a word that begins at frame τ and ends at frame t. Our local measure defines a neighbourhood around the analysed word $[w, \tau, t]$ by taking into account a fixed number of frames before and after the word. Thus, the total size of the neighbourhood V of a word w is defined by the sum of the length of w and the length of both past and future neighbourhoods. Figure 1 shows such a neighbourhood V of w with a past neighbourhood of length x and a future neighbourhood of length y.

The durations of the past and future neighbourhoods are independent. This allows us to use more data from the past (already processed, so available) without increasing the delay introduced by the future neighbourhood.

From the word graph generated by the recognition engine, we extracted the sub-graph corresponding to V and we com-

[1] http://www.cuedspeech.org

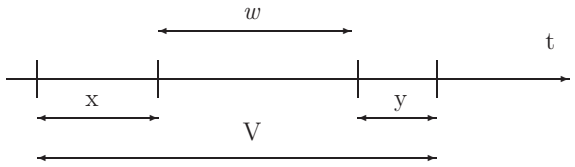

Figure 1: Neighbourhood V used for the computation of the confidence measure of w.

puted the estimation of the posterior probability of w. This estimation is obtained by the *forward-backward* method [5] computed at a word level and summarised by the following equations.

Let $\Phi([w,\tau,t])$ and $\Psi([w,\tau,t])$ denote respectively the *forward* and *backward* probabilities of the word $[w,\tau,t]$:

$$\Phi([w,\tau,t]) = p(o_\tau^t|w) \sum_{w_p} \sum_{\tau'} \Phi([w_p,\tau',\tau-1])\, p(w|w_p) \quad (1)$$

$$\Psi([w,\tau,t]) = p(o_\tau^t|w) \sum_{w_n} \sum_{t'} \Psi([w_n,t+1,t'])\, p(w_n|w) \quad (2)$$

In these equations, o_τ^t stands for the observation sequence between frames τ and t, w_p for a word preceding w and w_n for a word following w. The posterior probability of w is then computed as:

$$p([w,\tau,t]|o_1^T) = \frac{\Phi([w,\tau,t])\Psi([w,\tau,t])}{p(o_1^T)p(o_\tau^t|w)} \quad (3)$$

Knowing that the probability of the whole observation sequence $p(o_1^T)$ can be computed as:

$$p(o_1^T) = \sum_w \sum_\tau \Phi([w,\tau,T]) \quad (4)$$

In the extracted sub-graph associated with V, several occurrences of the analysed word may occur at similar but different temporal positions. The forward-backward method computes the posterior probability of each of these occurrences. Keeping only one occurrence under-estimates the true posterior probability of the word. In order to manage this problem, a flexibility factor η was introduced and the posterior probability of each occurrence of the analysed word satisfying several criteria according to η were added. Let d denote the length of the word w and $[\widetilde{w},\tilde{\tau},\tilde{t}]$ an occurrence of w that belongs to the sub-graph. We define the three following constraints:

- $\tau - \eta\, d \leq \tilde{\tau} \leq \tau + \eta\, d$
- $t - \eta\, d \leq \tilde{t} \leq t + \eta\, d$
- $(1-\eta)\, d \leq \tilde{d} \leq (1+\eta)\, d$

Let F stand for the set of the occurrences of w satisfying the previous constraints, the confidence $C([w,\tau,t])$ of w is defined by the following equation:

$$C([w,\tau,t]) = \sum_{[\widetilde{w},\tilde{\tau},\tilde{t}]\in F} p([\widetilde{w},\tilde{\tau},\tilde{t}]|o_b^e) \quad (5)$$

o_b^e denotes the observation sequence corresponding to the word sub-graph associated with $[w,\tau,t]$ and its neighbourhood V.

2.2 Speech Recogniton System

For this study, we used the large vocabulary speech recognition system ANTS [2]. This system is based on *Julius*, developed by researchers at the Kyoto University [3]. During the recognition process, Julius builds a word graph frame-synchronously from which the confidence value of a word is computed.

The acoustic parameterisation was based on MFCC by applying an MCR (Mean Cepstral Removal) normalisation.
The triphone models (HMM) were trained with HTK on a transcribed broadcast news corpus of around 40 hours extracted from a larger broadcast news corpus provided by the ESTER evaluation campaign in French in 2006 [1].
Language model was trained on 16 years of the french newspaper *Le Monde* and on manual transcription of broadcast news. The lexicon contains around 60,000 different words and the language model is composed of 19M of bigrams.

2.3 Assessment of our Local Measure

We assessed our confidence measure according to the Equal Error Rate (EER) criterion. Each recognised word was tagged as *accepted* or *rejected* by comparing its confidence value to a decision threshold. Then, two rates could be computed, false acceptance (FA) and false rejection (FR):

$$FA = \frac{number\ of\ incorrect\ words\ labelled\ as\ accepted}{num.\ of\ incorrect\ words}$$

$$FR = \frac{number\ of\ correct\ words\ labelled\ as\ rejected}{num.\ of\ correct\ words}$$

By varying the decision threshold value, we can find a particular operating point for which both acceptance and rejection rates are equal: the EER.

We evaluated the effect of the duration of path and future neighbourhoods on the accuracy of the local measure on a 1-hour development corpus of broadcast news [4]. Figure 2 shows the EER of the local measure according to the duration of the past neighbourhood. Three curves are plotted, corresponding to the duration of the future neighbourhood: 40, 60 or 84 frames (with a 10ms frameshift). The line plotted at 22% of EER corresponds to the performance obtained by a reference measure.
The reference measure is based on the estimation of the posterior probability computed on the whole signal once the speech recognition engine had processed the whole sentence. This reference measure is currently known to be one of the most accurate [6]. As our confidence measure had only a

partial knowledge of the audio signal, the performance obtained by this reference measure could be considered as a limit for our measure.

As expected, we can note that the larger the neighbourhoods are, the better the local measure performs. 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. The 84 frames correspond to the average duration of two consecutive words in the development corpus. Overall, we can observe that the local measure with a past and a future neighbourhood of 84 frames performed closely to the reference measure (EER are 23% and 22% respectively).
Therefore, we chose this local measure for the experiment described in the next section.

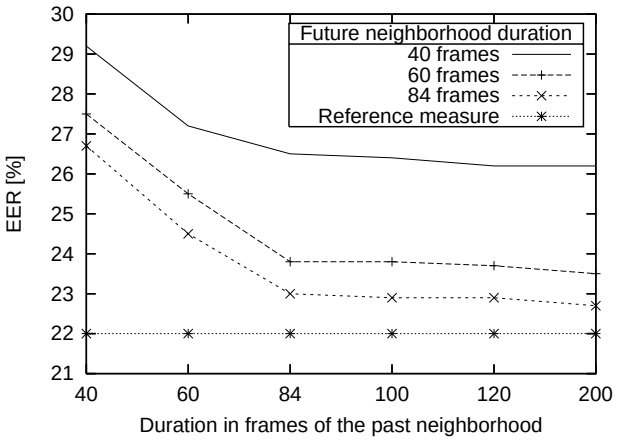

Figure 2: EER of the local confidence measure according to the duration in frames of both past and future neighbourhoods.

3. PRELIMINARY EXPERIMENT

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.

3.1 Modalities

The recognised words were transcribed according to the four following modalities:

Raw : the automatic transcription directly provided by the speech recognition system with no indication (all words in black colour);

Oracle : the automatic transcription whose misrecognised words were written in blue. A word is considered as misrecognised if it differs from the reference transcription;

Confidence : the automatic transcription whose words tagged as incorrect by our confidence measure were written in blue;

Phonetic : the automatic transcription whose words tagged as incorrect by our confidence measure were written using a simplified phonetic alphabet and in blue. Moreover, the successive incorrect words were concatenated before being transcribed.

Oracle modality allows us to assess the usefulness of a perfect confidence measure against a *Raw* transcription.

We introduced a *Phonetic* transcription of the words of low confidence because we think that, by this way, the reader can guess the good word by focusing on how the wrong word sounds rather than focusing on 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/, /u/ by /ou/, and /õ/ by /on/ because the grapheme " *u*" is pronounced /y/ in French etc.

The successive incorrect words were concatenated before being phonetically transcribed in order to remove the potential wrong word segmentations found by the recognition system. Table 1 shows the four modalities for a French sentence.

3.2 Experimental Protocol

In this experiment we wanted to assess the use of confidence measures according to two criteria: comprehension improvement and the user's feeling about the modalities. Our experimental protocol consisted in showing the automatic transcriptions of four texts to each subject participating in our experiment.

The four texts were initially designed as a reading level test for students entering secondary school (about 11 years old). The texts dealt with various topics: a fairy tale, a chronicle about a *Le Mans* car race, a story about a flight expedition, an investigation story about the theft of a computer.

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, neither for the speaker nor for the acoustic environment. It 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 by using the local confidence measure with a past and future neighbourhood of 0.84s. We then decided if the word is correct or not by comparing its confidence value with the decision threshold tuned on the development corpus (cf. 2.3). The low-confidence words (assumed as incorrect) were then highlighted in the cases of *Phonetic* and *Confidence* modalities.

We then compared the confidence value of a word with the decision threshold tuned on the development corpus (cf. 2.3) to decide if this word was correct or not. According to this decision, the low-confidence word were highlighted for the *Phonetic* and *Confidence* modalities.

The printed automatic transcriptions were submitted to 20 subjects with the following tasks to do in 15 minutes:

Table 1: Example of a word sequence presented according to the four modalities

Raw	Nous perdant de la hauteur une nourrice consigne continueront de percuter sommet[a]
Oracle	Nous **perdant** de la hauteur **une nourrice consigne continueront** de percuter sommet
Confidence	Nous perdant de la hauteur **une nourrice consigne continueront** de percuter sommet
Phonetic	Nous perdant de la hauteur **u_n_n_ou_r_i_s_k_on_s_i_gn_e_k_on_t_i_n_u_r_on** de percuter sommet
Original text	Nous perdons de la hauteur et nous risquons si nous continuons de percuter un sommet[b]

[a]We losing height a nurse deposit will going on hitting the peak
[b]We are losing height and we risk if we go on to hit a peak

Table 2: Example of a word sequence in English

Confidence	happy expense of actual **audits bury ants**
Phonetic	happy expense of actual **ao_d_ix_t_s_b_eh_r_iy_ae_n_t_s**
Original text	happy expense of actual experience

- to *restore* (guess) the original text of a part of the transcription (60 words on average);
- to answer a set of questions about the meaning of the rest of each text (about 10 questions per text);
- to answer a set of subjective questions about their feeling concerning the different modalities.

Each subject was given each of the four texts but with a different modality per text. In this preliminary experiment, the subjects participating in the test were hearing students (around 23 years old), but of course, they could not listen to the audio signal. Indeed, involving deaf students was difficult to manage for a first experimentation due to needing a sufficient number of participants as well as the availability of the subjects and their speech therapists.

3.3 Results

3.3.1 The Restored Text

First, we computed the word error rate (insertions, deletions and substitutions rates) for every *restored* text. We did not take into account spelling errors ; that is words that sound and mean in the same way but that are spelt differently. 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, aimer vs. aimÃl').
We did not take into account spelling errors because we wanted to assess the comprehension of the transcription and usually these kinds of errors have a very weak effect on the comprehension.

Table 3 shows this word error rate for each text and each modality. The ASR column gives the rate of the automatic recognition system.

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 we provided a visual clue in the word confidence either by the *Oracle* modality (exactly the wrong words) or by using a confidence measure (*Confidence* or *Phonetic* modality). It showed that the introduction of confidence measures helped the reader to correct an erroneous transcription.

On average, the *phonetic* modality gave the best results, but the difference is not significant. There may be two reasons to explain these performances. 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 is correlated to 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 1. The experiment was carried out in French, but in order to explain better, Table 2 shows an example in English. However this modality is maybe unusable for too young students or born deaf people.

3.3.2 Questions About the Meaning

Regarding the questions about the meaning of the text and which were essentially on the "incorrect" words, the results are difficult to interpret because the subjects gave the same answer to the majority of questions. For these questions the answers seem to be obvious or unguessable. It shows that it is difficult to find questions which satisfy all of the following constraints:

- questions must focus on words whose highlighting change according to the modality;
- answers should not be obvious or unguessable;
- the number of questions must be large enough to get significant results.

We can only conclude that highlighting the well-recognised words beacuse of the confidence measure did not disturb the subjects.

3.3.3 Subjective Questions

In their answers to the subjective questions, all the subjects expressed their preference for the *phonetic* modality because

Table 3: WER on the texts restored by the subjects according to the different modalities

Text	ASR [%]	Raw [%]	Oracle [%]	Confidence [%]	Phonetic [%]
Le Mans	18.8	10.4	**7.8**	11.0	9.0
Fairy tale	36.1	**14.2**	20.0	15.7	15.4
Flight expedition	47.4	47.4	40.7	36.5	**34.4**
PC's theft	20.5	17.9	**14.4**	16.9	16.4
Average	30.7	22.5	20.7	20.0	**18.8**

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. Indeed, people that have never heard may have a different relationship with some kind of phonetic memory.

Contrary to the results shown in Table 3, a majority of the subjects think that, writing the words hypothesized as wrong in blue colour, did not help them, and even disturbs them. But, the blue colour used for the phonetic modality did not disturb them.

For *confidence* and *oracle* modalities, several subjects pointed out that writing the low-confidence word in colour guide them towards a word having a similar root or meaning instead of a word that sounds in the same way.

4. CONCLUSIONS

We wanted to know if highlighting the words that are perhaps wrong can improve the comprehension of automatic transcription. First, we defined a local confidence measure that can be used in the framework of live streams. We then conducted an experiment to assess this hypothesis. Several conclusions and guidelines for future experiments can be drawn from this preliminary experiment.

We first showed that our local confidence measure can be useful in increasing the comprehension of recognised sentences.

Secondly, we evaluated two modalities to highlight low confidence words: only colouring them or also using a simplified phonetic alphabet. All the subjects preferred the *phonetic* modality. Above all, phonetic writing of potentially wrong words provides better results than just colouring them. However, this modality could be used by *becoming deaf* but not by people born deaf.

Furthermore, the decision threshold we used to tag the low confidence words as incorrect was set up according to the EER. This rate favours neither the false acceptance rate nor the false rejection rate. But the EER may not be the optimal operating point for this kind of task. Thus, other experiments with different operating points, based on perceptive criteria, should be carried out in order to assess the influence of the proportion of false acceptances and false rejections on the reader's understanding. Indeed, is it better to highlight a lot of words even if some of them are well recognised or to highlight few words even if it means that the user will not be aware of all possible misrecognised words ?

Highlighting the high-confidence words rather than the low-confidence words, should be included in a future experiment.

It could be also interesting to assess another modality than colouring as writing low-confidence words in smaller font.

This preliminary experiment showed that the *phonetic* modality achieved best results but it must be confirmed by a "live" experiment with scrolling text. But, the problem will be how to evaluate readers' understanding: the restoring task is impracticable and finding relevant questions is difficult.

This first experiment is interesting and encouraging because it shows that tools based on speech recognition and confidence measure 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 in the experiment with deaf students.

5. REFERENCES

[1] S. Galliano, E. Geoffrois, G. Gravier, J. Bonastre, D. Mostefa, and K. Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 315–320, 2006.

[2] I. Illina, D. Fohr, O. Mella, and C. Cerisara. The automatic news transcription system : Ants, some real time experiments. In *ICSLP*, pages 377–380, 2004.

[3] A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. In *EUROSPEECH, Aalborg*, pages 1691–1694, 2001.

[4] J. Razik, O. Mella, D. Fohr, and J. Haton. Frame-synchronous and local confidence measures for on-the-fly automatic speech recognition. In *INTERSPEECH, Brisbane*, 2008.

[5] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *ICASSP*, pages 875–878, 1997.

[6] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. SAP*, 9:288–298, 2001.

[7] B. Wrobel-Dautcourt, M.-O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low-cost stereovision based system for acquisition of visible articulatory data. In *AVSP*, pages 145–150, 2005.