



# CRF-based Combination of Contextual Features to Improve A Posteriori Word-level Confidence Measures

Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, Patrick Gros

IRISA (INRIA, University of Rennes 2, INSA, CNRS) - Rennes, France

firstname.lastname@irisa.fr

## Abstract

The paper addresses the issue of confidence measure reliability provided by automatic speech recognition systems for use in various spoken language processing applications. In this context, a conditional random field (CRF)-based combination of contextual features is proposed to improve word-level confidence measures. More precisely, the method consists in combining phonetic, lexical, linguistic and semantic features to enhance confidence measures, explicitly exploiting context information. The combination is performed using CRFs whose selected patterns enable to establish a precise diagnosis about the interest of individual and contextual features. Experiments, conducted on the French broadcast news corpus ESTER, demonstrate the added-value of the proposed CRF-based combination of contextual features, with significant improvement of the normalized cross entropy and of the equal error rate.

**Index Terms:** Confidence Measures, Feature Combination, Context, Spoken Language Processing

## 1. Introduction

Word-level confidence measures (CMs), indicating the reliability of a decision made by an automatic speech recognition (ASR) system, are key to many applications related to spoken language processing (SLP). Confidence measures may be used to detect errors [1], out-of-vocabulary (OOV) words [2, 3] or even to perform higher level SLP tasks (e.g., named entity recognition [4]). It has been already shown that taking CMs into account in SLP techniques improve performance. But, there is still a gap that, presumably, better CMs can bridge. This paper focuses on improving confidence measures with the specific aim of increasing their reliability for use by high level SLP. In this context, one of our major requirements is that the proposed method should be as independent as possible on a particular ASR system and should work in post-processing, i.e. mainly using the output information provided by the ASR system.

Among existing work on CM improvement [5], some propose to estimate the confidence of a word directly as its a posteriori probability, given acoustic observations [5, 6]. These methods provides fairly good results, especially when the probabilities are estimated from N-best hypothesis lists or word graphs [7], but fail to meet the requirements specified above. Indeed, they are dependent on a given ASR system and involve modifying its inner workings. Another way to compute confidence measures is to search for relevant clues, within the ASR output, that are sufficiently informative to distinguish correctly recognized words from possible recognition errors. These clues, called (predictor) features, are generally obtained during the decoding phase at either the acoustic, the language model, the syntactic or the semantic level. To improve performance,

these features are combined into a single measure that indicates the reliability of the recognized words. Many features have been studied [8, 5, 9] and several combination models have already been proposed. Although the results are still unclear in the literature [5], several studies show nevertheless the benefit of combining independent and complementary features [3]. Finally, few studies proposed to take into account the context of these features to improve confidence measures for large vocabulary speech recognition. It has however been shown that the contextual information could be beneficial for OOV word or error detection applications [3, 1, 10].

We propose a method combining individual and contextual features to improve word-level CMs for large vocabulary automatic transcripts. The approach consists in selecting a few features collected from various knowledge sources—phonetic, lexical, linguistic and semantic features—, as independent as possible of the ASR system and therefore easy to obtain. In order to increase the relevance of these features, information about the context (i.e. neighbor words) is also taken into account. Using the context is motivated by the fact that a recognition error on one word tends to impact the following words due to the N-gram language model. Finally, so as to efficiently combine features with their contextual information, a machine learning method based on conditional random fields (CRF) is investigated. CRF, based on a discriminant model able to integrate different kinds of features, are also characterized by their ability to manage sequential data and thus contextual information. Finally, CRF have the advantage of producing patterns enabling to establish a precise diagnosis about the interest of individual and contextual features.

Section 2 details the proposed approach. After describing the experimental setup in Section 3, Section 4 reports experiments and results regarding the contribution of each base and contextual feature, the contribution of their combination and the most relevant patterns obtained from CRF training. Finally, conclusions are given in Section 5.

## 2. A word-level confidence measure combining contextual features

The method proposed to improve a posteriori word-level confidence measures by combining contextual features consists in 3 steps: selecting base features, building contextual features by enriching the base features with their context and combining all features with a CRF-based classifier. These steps are now described in more details.

### 2.1. Base features

For each word in the ASR system hypothesis, a few relevant features have been selected. The following features, classified

feature	w	pos	#ph	lmbb	dur	ne	conf
# classes	64937	144	16	10	6	3	3

Table 1: Number of classes for each extracted feature

by category, are considered:

*A posteriori*. confidence class (**conf**)—Our ASR system already provides an a posteriori CM [6] based on N-best lists that will be used as a baseline in the experiments. The confidence class is obtained by discretizing this CM into 3 classes.

*Lexical*. word (**w**)—Although this feature has no generalization ability, it will be useful to detect systematic ASR errors.

*Morphosyntactic*. part-of-speech (**pos**)—Transcripts are tagged with a set of 144 POS classes containing general morphosyntactic classes as well as very frequent words. This feature enables to know the a priori error distribution for each POS class. Although simple, it is not so much used for computing confidence measures.

*Linguistic*. language model back-off behavior (**lmbb**)—In a word sequence, for each word, the language model (LM) back-off behavior is the degree  $n$  of the largest current  $n$ -gram belonging to the chosen LM. It has been shown to be correlated with errors [11, 3]. The 4-gram language model from our ASR system is used here but any other could have been chosen independently from the ASR system. This feature is composed of 4 main classes ('11', '12', '13', '14') and 6 other specific classes ('11', '21', '22', '31', '32', '33') that represent the different cases in beginning of sequence. For a class 'xy',  $x$  and  $y$  represent respectively the position in the sequence (the first 3 positions '1', '2', '3' and the next ones '1' inside the sequence) and the LM degree  $n$  as defined previously ( $n$  from 1 to 4).

*Phonetic*. duration class (**dur**) and number of phonemes (**#ph**)—Many observations points out that word length can help in predicting correct words and errors (e.g. OOV words tend to be misrecognized as a sequence of short words).

*Semantic*. named entity detection class (**ne**)—The main difficulty for SLP engines to process automatic transcripts is the presence of recognition errors. Whereas confidence measures can help in tackling the problem, SLP techniques can also, by feedback, be useful to estimate CMs. The idea is to use the agreement between similar SLP engines as a semantic descriptor. We propose to use three different named entity (NE) taggers [12] to design a semantic feature 'ne' composed of 3 classes: 'NE' (if the 3 taggers recognize the same NE), 'NaNE' (if the 3 taggers recognize that the word is Not a NE), '?' (else in ambiguous cases). We believe that these ambiguous cases are induced by recognition errors, especially on semantically rich words. Note that apart from named entities, other features could be designed in the same way using different SLP applications to identify semantic ambiguities.

Since CRF consider symbolic features, continuous features are discretized to fulfill this condition. To do so, a C4.5 decision tree was used to minimize the entropy of each class. Table 1 shows the number of classes for each extracted feature.

## 2.2. Contextual features

It is a well-known fact that an ASR error often impacts the following words. Using context to enrich base features have shown promising results [1, 3] that led us to consider various shapes of context. We propose to build new features, called "*contextual features*", from a base feature by exploiting the context. This process includes 3 steps: defining the context, defining  $n$ -gram patterns, and selecting the best patterns for each base feature.

-3	-2	-1	0	1	2	3
le	tour	de	france	troisième	étape	remportée

Table 2: Example of the word feature 'france' in its context

Firstly, the *context* of a feature at a current position in a sequence is composed of the  $s$  neighbors on both sides. Empirically, we chose  $s = 3$  as we observed that a longer context size was not necessary. Secondly, we define  $n$ -gram *patterns* that associate  $n$  different positions of the context to create new features, where  $n \in [1, 7]$  because  $s = 3$ . For instance, in Table 2, the 3-gram pattern '-2/-1/0' for feature  $w$  associates the words 'tour', 'de' and 'france' to create the new feature 'tour/de/france'. Finally, we *select* the best patterns for each base feature by evaluating, as defined in Section 3, all the possible combinations of  $n$ -gram patterns.

## 2.3. CRF-based combination and confidence measure

We consider the two classes problem of labeling automatic transcripts with the labels 'correct' or 'erroneous'. Many algorithms have been investigated to label sequential data (SVM, boosting, decision trees, hidden Markov models, CRF, etc.). In our context, we need a machine learning algorithm able to deal with descriptors coming from different sources of knowledge, which is the prerogative of discriminant algorithms. Since errors are highly dependent one on another, an algorithm able to make a global decision on the sequence is required. We use CRF<sup>1</sup> a probabilistic model dedicated to labeling sequential data [13]. Like classification models, CRFs accommodate many statistically correlated features as input and use discriminative training. But unlike many of them (SVM, perceptron, etc.) which views the sequential labeling problem as a set of independent decisions, CRFs compute the probability of a sequence of labels given a sequence of observations, thus taking a global decision on the sequence. Moreover, CRFs enable to estimate easily a marginal probability of each decision in the sequence, the marginal probability of the label 'correct' being used as the CM. Another interest is that the CRF classifier weights each feature during the training stage, making it possible to interpret the most relevant "rules" for predicting correct or erroneous words.

## 3. Experimental setup

Experiments are carried out with a large vocabulary radio broadcast transcription system, exhibiting error rates around 20% on broadcast news data. Hidden Markov phonetic models were trained using approximately 200h of speech material. A 4-gram language model was obtained from about 500 million words mostly coming from French-speaking newspapers. Confidence measures are provided based on posterior probabilities combining acoustic, language model and POS scores as in [6]. Results are reported on the corpus from the French evaluation campaign ESTER 2 [14], consisting of 12 hours of French radio broadcasts. The ESTER 2 corpus is officially divided into a development and a test set of 6h each, for which word error rates of respectively 31.7% and 25.2% were achieved. We used the development set to train the CRF classifier and the test set to evaluate performance. Two standard evaluation metrics are used, the equal error rate (EER) and the normalised cross entropy (NCE), in addition to detection error trade-off curves

<sup>1</sup>CRF++ (<http://crfpp.sourceforge.net/>) is used in this work.

feature	base			contextual		
	EER	NCE	#feat	EER	NCE	#feat
<b>baseline</b>	<b>30.81</b>	<b>-0.074</b>	.	<b>30.81</b>	<b>-0.074</b>	.
#ph	46.82	0.012	28	44.20	0.024	334
ne	46.71	0.017	10	45.02	0.018	86
dur	44.99	0.019	14	43.05	0.030	180
w	42.03	0.029	180	41.06	0.039	692
pos	41.03	0.036	172	37.24	0.069	1138
lmbb	38.81	0.062	24	35.31	0.094	302
<b>conf</b>	<b>30.64</b>	<b>0.168</b>	10	<b>29.71</b>	<b>0.176</b>	92

Table 3: Results for each base and contextual feature in equal error rate (EER), normalised cross entropy (NCE) and number of generated features (#feat)

plotting true error detection rate vs. false alarm rate.

## 4. Experiments and results

We first study the contribution of each individual feature before evaluating the combination and analyzing the most relevant contextual patterns.

### 4.1. Contribution of each base feature

Practically, for each feature, the pattern ‘0’ is used to train the CRF-based classifier and the resulting model is evaluated on the test set. Results are reported in Table 3.

The best result is obtained with the feature *conf* with an EER equivalent to that of the baseline (30.64 vs. 30.81). This confirms that the confidence class and the a posteriori CM bear the same information and that discretization does not affect the quality of the confidence measure. However, NCE does not support this obvious interpretation which explains why NCE is criticized [7]. Other features are far below the baseline. Indeed, each feature has been designed to add a specific piece of information to the global CM (*conf*). As expected, additional features do not individually achieve satisfactory results. We can however notice that linguistic and morphosyntactic features are better than the lexical one.

### 4.2. Contribution of each contextual feature

Contextual features as described in Section 2.2 are studied and we evaluate the contribution of each contextual feature separately.

Table 4 reports, for each base feature, the best pattern combination and shows that, in most cases, using only the previous and next positions as context is enough to enrich the current position. This supports the observation in [3] that “a window larger than 3 words leads to worse results”. However, a noticeable point is that the linguistic (*lmbb*) and semantic (*ne*) features resulted in better performance with a larger context (patterns ‘1/2/3’, ‘-2/-1/0’ and ‘1/2’). This means that the relevant contextual information are different depending on the nature of the selected features.

Table 3 reports the performances obtained for each contextual feature. One can notice that, with no exception, using the context always results in increased performance. An interesting result is that the confidence class ‘*conf*’ enriched by the context ‘-1/0/1’ yields better results than the baseline (29.71 vs. 30.81). This means that a posteriori CMs can be improved by using their own context.

Finally, this set of results corroborates the fact that errors are strongly dependant, a fact known as *error propagation*,

feature	pattern combination
<i>conf</i>	-1/0/1
<i>lmbb</i>	-1/0/1, 1/2/3
<i>pos</i>	-1, 0, 1, -1/0, 0/1
<i>w</i>	-1, 0, 1, -1/0, 0/1, -1/1
<i>dur</i>	-1/0/1
<i>ne</i>	-2/-1/0, 1/2
<i>#ph</i>	-1/0, 0/1

Table 4: Results for the best patterns combination found to enrich each base feature with its context (‘0’ represents the current position, ‘-n’ the n-th previous feature in the sequence, ‘n’ the n-th next feature in the sequence and ‘a/b’ the pattern associating the feature at position ‘a’ and the feature at position ‘b’)

feature	base			contextual		
	EER	NCE	#feat	EER	NCE	#feat
<b>baseline</b>	<b>30.81</b>	<b>-0.074</b>	.	<b>30.81</b>	<b>-0.074</b>	.
<b>conf</b>	30.64	0.168	10	29.71	0.176	92
<b>+lmbb</b>	28.99	0.179	30	26.87	0.206	390
<b>+pos</b>	28.52	0.188	198	25.54	0.230	1524
<b>+dur</b>	27.95	0.198	208	24.90	0.239	1700
<b>+#ph</b>	26.93	0.212	232	24.39	0.249	2030
<b>+ne</b>	26.78	0.215	238	24.39	0.239	2112
<b>+w</b>	<b>26.74</b>	<b>0.216</b>	414	<b>24.22</b>	<b>0.251</b>	2800
- <i>conf</i>	33.49	0.115	408	29.22	0.167	2712
- <i>lmbb</i>	37.33	0.075	388	34.45	0.098	2414

Table 5: Results for the combination of base and contextual features in equal error rate (EER), normalised cross entropy (NCE) and number of generated features (#feat)

since contextual information increase word CM estimation for all features.

### 4.3. Combination of features

Finally, we evaluate the combination of features using either base features or contextual ones. Practically, features are combined iteratively in the best way (i.e. the best combination is chosen at each step) until there is no more increase in performance.

Results are reported in Table 5 for the successive combinations of the following features (in order of importance): *conf*, *lmbb*, *pos*, *dur*, *#ph*, *ne* and *w*. The final combination is also evaluated without selecting the features *conf* and *lmbb*.

Experimental results show that the features are complementary with the a posteriori CMs and helps to improve it. Indeed, a significant improvement in EER and NCE is obtained with respect to the baseline (30.8/-0.074) using either base features (26.7/0.216) or contextual features (24.2/0.251). However, the lexical (*w*) and semantic (*ne*) features have moderate contributions because words are too specific and redundant with the POS feature while only a small, yet semantically meaningful, number of words correspond to named entities. Surprisingly, phonetic features (word duration and number of phonemes) are complementary. Last but not least, combining features without considering the confidence class still gives slightly better performance than the baseline in the contextual case (29.44/0.166 vs. 30.81/-0.074). This means that it is possible to compute a CRF-based confidence measures as reliable as a posteriori ones, without relying on the ASR system CMs.

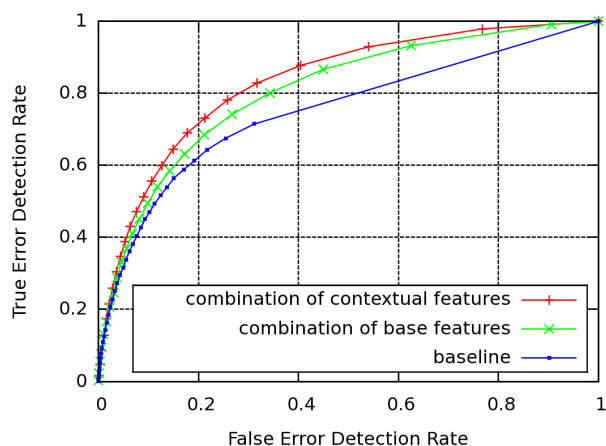


Figure 1: ROC curve for the baseline, the best combination of base features and the best combination of contextual features

#### 4.4. Pattern analysis

After training, the CRF-based classifier yields a model of weighted patterns used to predict the word correctness at each position of the sequence. We propose to analyse the most relevant patterns to assess the conditions in which erroneous and correct words appear.

The CRF classifier clearly shows that the lmbb feature is relevant to predict errors. The best pattern to predict correct words is 'I4/(I4)/I4', i.e. no back-off used for the previous, current and following word. Conversely, the best pattern to predict errors is 'I2/(I1)/I1' where strong back-off is used for every word. The morphosyntactic feature is also quite relevant: CRF also confirms that some part-of-speech categories are more sensitive to errors than others, in particular errors due to morphological variations. For instance, words in the category 'past participle in the singular feminine form' are often erroneous because their morphological form varies according to gender and number contrary to their pronunciation (homophones). The most relevant patterns for the lexical feature (w) enable us to find that very specific words are systematically correct or erroneous. For instance, "président" and "aujourd'hui" are often well-recognized while "là" and "elle" are misrecognized in most cases. Sequence of short words (less than 3 phonemes) are also an important clue to detect errors where, e.g., the OOV word "cancellara" (proper name) is recognized as the sequence of four short words "quand c' est lara". Regarding the semantic feature, the CRF classifier indicates the cases when the 3 NE taggers disagree are source of errors because the context is ambiguous. When all the named entity taggers recognize the same entity, the word is in most cases correct.

### 5. Conclusion

This work tackles the difficult problem of estimating word-level confidence measures for transcripts obtained from a large vocabulary ASR system. A CRF-based combination of individual and contextual features to improve word-level confidence measures have been proposed and evaluated. Experiments on a French radio broadcasts corpus exhibit several interesting results. The first one is that we have confirmed that feature combination is beneficial to improve confidence scores, thus demonstrating the interest of combining features collected from vari-

ous knowledge sources, e.g., lexical, phonetic, morphosyntactic, semantic and linguistic features. Another benefit is that the selected features can be obtained independently of any ASR system, either directly from the output provided by systems, or from external tools such as natural language processing ones. This point is crucial in the context of this work whose main focus is to obtain more reliable confidence measures for use by high level spoken language processing techniques, independently from a particular ASR system. Experiments have also shown the added value of the contextual information associated with features: the gain is significant and systematic for all proposed features. A detailed analysis of contexts demonstrated that the best context is limited to the immediate neighbors and that the relevant contextual information may varied according to the features (e.g., the LMBB features). Finally, results also indicate the efficiency of the proposed approach to combine individual feature and contextual information. The CRF-based method enables to efficiently combine all kinds of features and to produce patterns describing the contribution of selected features.

### 6. References

- [1] G. Skantze and J. Edlund, "Early error detection on word level," in *COST278 and ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction*, 2004.
- [2] H. Sun, G. Zhang, F. Zheng, and M. Xu, "Using word confidence measure for OOV words detection in a spontaneous spoken dialog system," in *Eighth European inproceedings on Speech Communication and Technology*, 2003.
- [3] B. Lecouteux, G. Linarès, and B. Favre, "Combined low level and high level features for Out-Of-Vocabulary Word detection," in *Interspeech*, 2009.
- [4] B. Favre, F. Béchet, and P. Nocéra, "Robust named entity extraction from large spoken archives," in *HLT-EMNLP*, 2005, pp. 491–498.
- [5] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.
- [6] S. Huet, G. Gravier, and P. Sébillot, "Morpho-syntactic post-processing of n-best lists for improved french automatic speech recognition," *Computer Speech and Language*, no. 24, pp. 663–684, 2010.
- [7] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.
- [8] D. Yu and L. Deng, "Semantic confidence calibration for spoken dialog applications," in *ICASSP*, 2010, pp. 4450–4453.
- [9] T. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech & Language*, vol. 16, no. 1, pp. 49–67, 2002.
- [10] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition," in *ICASSP*, 1997, pp. 875–878.
- [11] J. Maclair, Y. Estève, S. Petit-Renaud, and P. Deléglise, "Automatic detection of well recognized words in automatic speech transcription," in *LREC*, 2006.
- [12] C. Raymond and J. Fayolle, "Reconnaissance robuste d'entités nommées sur de la parole transcrit automatiquement," in *TALN*, Montréal, Canada, July 2010.
- [13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML'01*, 2001, pp. 282–289.
- [14] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Interspeech*, 2009, pp. 2583–2586.