Fuzzy integrals for the aggregation of confidence measures in speech recognition

Julie Mauclair, Laurent Wendling, David Janiszek

Fuzzy Integrals For The Aggregation Of Confidence Measures In Speech Recognition

Julie Mauclair, Laurent Wendling, David Janiszek
LIPADE, University of Paris Descartes
Paris, France
Email: {julie.mauclair,laurent.wendling,david.janiszek}@parisdescartes.fr

***Abstract*—This paper presents a study on merging confidence measures using fuzzy logic. Instead of the previous approaches using the notion of probability, we propose to observe the uncertainty of the recognition hypotheses and the notion of possibility thanks to fuzzy reasoning. Four different confidence measures are developed, coming from different parts of a speech recognizer. Various merging methods are studied to improve the performance of the confidence measures. The methods are evaluated in terms of Confidence Error Rate (CER) and in terms of their Detection Error Tradeoff (DET) curves on a French broadcast news corpus. They are compared to some fuzzy logic aggregation techniques among which the technique based on the Choquet Integral yields to a significant improvement in terms of CER.**

***Keywords*—fuzzy logic, confidence measures, speech recognition, feature aggregation**

I. Introduction

Confidence measures are used in various applications of speech processing [1]. When used with a threshold, a confidence score associated to an hypothesis can lead to the acceptation or rejection of that hypothesis [2]. Confidence measures can also be used to select hypothesis that will increase the amount of training data needed for acoustic models [3]. They can also guide the confirmation strategy in dialog systems [4] and can be used for language identification [5]. In speech recognition, confidence measures give an estimation about the correctness of a given hypothesis. To provide such estimations, many parts of the speech recognition system can be helpful as they can provide an indication about the reliability of the system.

One way to take advantage of the quality of each measure is to find the most performant combination method. Designers often choose between the techniques available and the techniques they know best by simply making an educated guess. A tempting approach is to combine several decision rules, based on various representations and classification schemes, instead of electing only one rule. The expected outcome of an aggregation is a more robust final decision, that will improve the classification ability of a single confidence measure. Well-known operators applying this scheme include the quasi-arithmetic means, the weighted minimum and maximum, and the ordered weighted averaging [6]. Other techniques such as Support Vector Machines (SVM), neural networks and Gaussian Mixture Models (GMM), can also be used to merge scores [7].

However, none of these families of operators can take the possible interactions between the constituents of the aggregation into account. The Choquet integral is then considered because it makes it possible for such interactions to generalize many aggregation operators by choosing specific fuzzy measures such as weighted arithmetic means, ordered weighted averages, order statistics, and median [8], [9], [10]. Fuzzy integrals, and the Choquet integral in particular, have been successfully used as fusion operators in various applications of pattern recognition [11] . The aim of the paper is then to explore how fuzzy integrals and the selection of measures can improve the confidence annotation of recognition hypotheses in speech processing.

In this paper, we will first present the different confidence measures developed for our study. The third section will describe the various aggregation methods which are taken as baselines. In section IV, the two fuzzy logic systems which are used for the experiments are presented. In section V, the corpora used will be detailed along with the evaluation metrics. Some results for each merging method will also be described. Finally, section VI presents how the measures can be selected to improve the results.

II. Confidence measures

Let us consider a set of N recognized words $\{w_1, \ldots, w_N\}$. Each word w can be associated with a confidence measure $m(w)$ defined with the properties:

Property 1:
The measure should be in the usual domain $[0, 1]$ and the measure should be interpretable as the probability that the word w is correct.

Property 2:
As a consequence of the latter property, the correct recognition rate on the emitted words is given by the following approximation:

$$\frac{1}{N}\sum_{i=1}^{N} m(w_i) \approx \text{CWRR} \quad (1)$$

where CWRR is the Correct Word Retained Rate (deletions are not taken into consideration).

A. Acoustic measure (AC)

The acoustic measure is based on the comparison of the acoustic likelihood provided by the speech recognition system for a given hypothesis and the one provided by an unconstrained phone loop model defined in [12]:

$$m^*_{ac}(w) = \frac{1}{N_f(w)}[\log P(Y|\lambda_C) - \log P(Y|\lambda_L)] \quad (2)$$

where w is the recognized word with N_f frames, Y is the sequence of acoustic observations, $P(Y|\lambda_C)$ is the acoustic score given by the recognizer model, and $P(Y|\lambda_L)$ is the acoustic score given by an unconstrained phone loop.
This measure is then normalized using a sigmoid-like transformation.

$$m_{ac}(w) = \frac{\exp\left(\frac{m^*_{ac}(w)-\mu}{\sigma}\right) + a}{\exp\left(\frac{m^*_{ac}(w)-\mu}{\sigma}\right) + 1} \quad (3)$$

where μ, σ are the mean and the standard deviation of the initial acoustic measure on a training corpus and $a = 2 * \text{CWRR} - 1$ (for an approximation of the second property defined at the begining of this section).
The value of this measure for a word w will be noted $m_{AC}(w)$.

B. Language Model Back-off Based measure (LMBB)

This measure was previously defined in [12]. It uses the back-off behavior of the language model (LM) to define the confidence score that is associated to a word.
A given word recognized with a given left context is associated with the highest order of n-grams seen in the training corpus concerning this word and this context. For a quadrigram LM, the order can be as high as 4, or it can be 0 if out-of-vocabulary words can be processed.
For example, if the sequence of words 'it is the ninth time' is recognized using a quadrigram model and if the quadrigram [is the ninth time] was observed in the training corpus, 'time' will be associated with the order 4. But if this quadrigram has never been observed, whereas the trigram [the ninth time] has, 'time' will be associated with the order 3, and so on, all the way down to order 1 (or 0 if out-of-vocabulary words can be processed).

It is well known that an error occurring in a word has an impact on the correctness of the words located in the immediate context of this erroneous word. According to this, and assuming that the LM back-off behavior is an acceptable criterion to predict the correctness of a word, our measure considers the LM back-off behavior of the left and right neighbors of a word, in addition to the highest order of n-grams this word is associated with on a training corpus.

So, each recognized word is associated with a triplet corresponding to its class.

In order to reduce the number of classes, each word of a recognized hypothesis is finally associated with a three component label:

1^{st} symbol: -, =, or + when the order of its left neighbor is respectively lower than, equal to, or higher than the order of the considered word,

2^{nd} symbol: the highest order of observed n-grams associated to the given word

3^{rd} symbol: -, =, or + when the order of its right neighbor is respectively lower than, equal to, or higher than the one of the considered word.

By comparing a set of automatic transcriptions with words labeled with these triplets, to a manual transcription of the same set of sentences, the misrecognition rate is computed [1] for each triplet on a training set. This rate is the number of substitutions and insertions divided by the number of recognized words. The misrecognition rate for a given class will later be used as the confidence measure for words labeled with that triplet when processing test data. Indeed, in [12], we showed that there is a correlation between the triplets and the error rate those words are associated with.
The value of this measure for a word w will be noted $m_{LMBB}(w)$.

C. Word posterior probability measure (WP)

Word posterior probabilities can be computed from N-best lists, word-lattices [13] or confusion networks [1], [14]. Roughly, the word posterior probability is the ratio of the *a posteriori* probability of a word and the sum of the *a posteriori* probabilities of all its alternatives. These *a posteriori* probabilities result from a combination of values given by acoustic and language models. Thus, word posteriors can be seen as a summarization of acoustic scores, linguistic scores and the search space topology. In N-best lists, the word posterior probability of a word is approximated with the ratio of the sum of the *a posteriori* probabilities of the occurrences of this word in the N hypotheses in a given position, and the sum of all the *a posteriori* probabilities of occurrences of words in this same position, including occurrences of the given word. In both word lattice and confusion network based approaches, the word posterior probability can be seen as a generalization of the N-best approach, where word-segmentations and search space depth are better considered.

Unfortunately, this measure is affected by pruning heuristics reducing the size of pruned word lattices generated during the recognition process. In practice, the use of this measure can therefore be biased. To overcome this problem, a decision tree can be trained to transform the posterior probabilities into better confidence scores [13].
In this paper, we use a confusion network based approach, directly derived from [15] to compute word posteriors. The value of this measure for a word w will be noted $m_{WP}(w)$.

D. Word posterior probability mapped measure (MAP)

An analysis showed the tendancy of the WP confidence measure to overestimate the probabilities of correct recognition as it was also described in [14]. This is due to the fact that

[1] In this paper, we use 'misrecognition rate' to refer to insertions and substitutions errors only, whereas 'word error rate' refers to the common metric including insertions, substitutions *and deletions*; this distinction is needed when studying recognized words only. The misrecognition rate equals 1-CWRR, the Correct Word Retained Rate previoulsy mentioned.

the lattices used as the basis for the posterior estimation only represent part of the posterior distribution and a significant amount of the probability mass "is missing" as we have explained in the previous section. Therefore, a piecewise linear mapping function is used in order to make this measure more discriminant. The coefficients are computed on a specific training set used for the parameters of the confidence measures and led to a new measure called MAP.
The value of this measure for a word w will be noted $m_{MAP}(w)$.

By aggregating the word posterior probability with other confidence measures which are not affected by the search space size, the performance of the word posterior probability should be improved.

III. Aggregation methods : baselines

We have four measures that come from different parts of the speech recognizer. In order to get the best of each measure, several merging techniques are studied. Various classifiers and aggregation techniques have been computed in order to compare their performance against fuzzy systems.

A. Arithmectic mean

One combination that can be a baseline is a simple arithmetic mean between all the features.

B. Linear least square regression

To take into account the quality of each measure, a simple linear interpolation can be used to fit a predictive model of the word correctness as in the equation:

$$y = \sum_{i=1}^{n} \alpha_i x_i + \alpha_0 \tag{4}$$

The method of the least squares is used to estimate the values of the interpolation parameters on the MTrain data set (see section V-C).
Various combination using the four measures or only three of them are tried.

C. Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron with two hidden layers is trained with backpropagation. Each measure gets one input node and there is one ouput node.

IV. Fuzzy logic systems

In general, uncertainty in the probability theory is seen in terms of occurrences of known facts. In a speech recognition task, what is known is whether the recognition hypothesis is correct or not. Probability is useful when dealing with serial events that require an enumeration notion of uncertainty but is not very useful when the uncertainty is about the degree of accomplishment of a known situation [16]. Indeed, when applying confidence measures to speech recognition, what we want to know is mainly about the degree of correctness that is associated with the recognition hypotheses. Fuzzy logic systems are therefore relevant for this task because they use the notion of uncertainty from the point of view of "possibility".

Fuzzy logic systems use a set of fuzzy rules to map a number of fuzzy inputs to fuzzy outputs. They allow the classification of fuzzy variables thanks to "if ... then" rules [17].

A. Possibility measures

In [18], the authors express the transformation between probability and possibility by:
Let C be the set composed of $c_1, c_2, ..., c_n$, $p_i = P(c_i)$ be the probability measures of each element of the set C where $p_1 \geq p_2 \geq ... \geq p_n$, and $\pi_i = \Pi(c_i)$ be the possibility measures on the set C, then the optimal solution is :

$$\forall i = 1, n, \pi_i = \sum_{j=i}^{n} p_j \tag{5}$$

This transformation is then used to compute the possibility measures from the scores given by the four experts. Various aggregation techniques can be used to get the final confidence score. In this article, the normalised conjonctive, the disjonctive and the adaptative fusions are computed.

B. Fuzzy Integrals

Let us denote by $X = \{D_1, \ldots, D_n\}$ the set of n decision rules (DR), and $\mathcal{P}$ the power set of X, i.e. the set of all subsets of X.
Definition 1 *A* fuzzy measure *or* capacity, μ, *defined on X is a set function* $\mu : \mathcal{P}(X) \longrightarrow [0, 1]$, *verifying the axioms:*

$$\mu(\emptyset) = 0, \mu(X) = 1 \tag{6}$$

and

$$A \subseteq B \Longrightarrow \mu(A) \leqslant \mu(B) \tag{7}$$

Fuzzy measures generalize additive measures, by replacing the additivity axiom by a weaker one (monotonicity). Fuzzy measures embed particular cases including probability measure, possibility and necessity measures, or belief and plausibility functions.
In our context of decision rule fusion, $\mu(A)$ represents the importance, or the degree of trust in the decision provided by the subset A of DRs. The next step in building a final decision, is to combine the partial confidence degree according to each DR into a global confidence degree, taking those weights into account.

1) The Choquet Integral: The Choquet integral was first introduced in the capacity theory [19], [20].

Definition 2 *Let* μ *be a fuzzy measure on X. The* discrete Choquet *integral of* $\vec{\phi} = [\phi_1, \ldots, \phi_n]^t$ *with respect to* μ, *noted* $\mathcal{C}_\mu(\vec{\phi})$, *is defined by:*

$$\mathcal{C}_\mu(\vec{\phi}) = \sum_{j=1}^{n} \phi_{(j)}[\mu(A_{(j)}) - \mu(A_{(j+1)})] \tag{8}$$

where $(.)$ *is a permutation on the source indexes, such as* $(i) \leq (j) \Rightarrow \phi(i) \leq \phi(j)$*. Also,* $A_{(j)} = \{(j), \ldots, (n)\}$ *represents the* $[j..n]$ *associated criteria in increasing order and* $A_{(n+1)} = \emptyset$*.*

Determining the Fuzzy Measure

There are several methods to determine the most adequate fuzzy measure to be used for a given application and the most straightforward learning approach is based on optimization techniques. The aim is to find the fuzzy measure that best minimizes a criterion on the training set, such has the square error. Considering $(x^k, y^k), k = 1, \ldots, l$, l learning samples where $x^k = [x_1^k, \ldots, x_n^k]^t$ is a n-dimensional vector, and y^k the expected global evaluation of object k, the fuzzy measure can be determined by minimizing [8]:

$$E^2 = \sum_{k=1}^{l} (\mathcal{C}_\mu(x_1^k, \ldots, x_n^k) - y^k)^2 \tag{9}$$

This criterion can be put under a quadratic program form and solved by the Lemke method. Nevertheless the method requires at least $n!/[(n/2)!]^2$ learning samples. When little data is available, matrices may be ill-conditioned, causing a bad behaviour of the algorithm. To cope with the above problems, "heuristic" algorithms were developed.
To our knowledge, the algorithm providing the best approximation was proposed by Grabisch in [11]. It assumes that in the absence of any information, the most reasonable way to aggregate the partial matching degrees is to compute the arithmetic mean on all the inputs. This aggregation using Choquet Integral will be later noted m_{CI}.

The fuzzy measures used in the Choquet Integral method can also be computed thanks to possibility measures (see section IV-A). This aggregation will be later noted m_{CF}.

2) The Sugeno Integral: A Sugeno-type fuzzy inference system has already been proved to be a good classifier [21]. This system was used in [22] with good results on a Spanish speech database. This database was collected via land telephone lines (sampled at 8kHz). On the continous speech database (consisting of 9405 words) the Equal Error Rate went from 22.85 with an MLP down to 22.05 when using a Sugeno-type fuzzy inference system.

For our experiments, the number of input variables corresponds to the four confidence measures and the ouput will be the value of the resulting confidence measure.

Determining the Fuzzy Measure

In this paper, the same kind of Sugeno Integral that was used in [22] is computed, using grid partitioning, and yielded to 16 rules. The generalized bell membership function is used for the fuzzification of the scores given by the speech recognizer. This method of aggregation, using the Sugeno Integral, will later be called m_{SI}.

3) Behavioural Analysis of the Aggregation:

Important Index

The importance index is based on the definition proposed by Shapley in game theory [23]. Its application in the context of fuzzy measures was made by Murofushi and Soneda [24]. It is defined for a fuzzy measure μ and a rule i as:

$$\sigma(\mu, i) = \frac{1}{n} \sum_{t=0}^{n-1} \frac{1}{\binom{n-1}{t}} \sum_{\substack{T \subseteq X \setminus i \\ |T|=t}} [\mu(T \cup i) - \mu(T)] \tag{10}$$

It can be interpreted as an average value of the marginal contribution $\mu(T \cup i) - \mu(T)$ of the decision rule i alone in all combinations.
The measure using the importance index will be noted m_{II}.

Interaction Index

Another useful index in order to apprehend the degree of interaction between decision rules was also introduced by Murofushi and Soneda [24]. The interaction value between the rules i and j, to the presence of elements of the combination $T \subseteq X \setminus ij$ is computed as:

$$(\Delta_{ij}\mu)(T) = \mu(T \cup ij) + \mu(T) - \mu(T \cup i) - \mu(T \cup j) \tag{11}$$

Extending this criteria on every subsets of $T \subseteq X \setminus ij$, an evaluation of the interaction between DRs i and j is :

$$I(\mu, ij) = \sum_{T \subseteq X \setminus ij} \frac{(n-t-2)!t!}{(n-1)!} (\Delta_{ij\mu})(T) \tag{12}$$

This index gives an idea of the complementarity or competitiveness that exist beween measures. A positive interaction between two rules means that the decision rule is discriminative power is increased when used with th second decision rule. Thus, this index will later by used to select measures in order to obtain an aggregation of complementary measures.

V. Experiments

Experiments have been carried out on the ESTER corpus. ESTER is an evaluation campaign of French broadcast news transcription systems which started in 2003 and completed in January 2005 [25]. The recognition system used for those experiments is the LIUM (Laboratoire d'Informatique de l'Université du Maine) system [26]. It is based on the CMU Sphinx 3.3 decoder. Some features were added to the Sphinx decoder, such as acoustic models adaptation using Speaker Adaptative Training (SAT) method or word-lattice rescoring to get a quadrigram language model. More details about the system can be found in [26]. The LIUM system reached the second position of the campaign with an official 23.6% word error rate.

A. Resources

The vocabulary used by the LIUM system contained about 65K words. Acoustic and language models were trained using the official training corpus of the ESTER evaluation campagn. Trigram language model was used during the first two passes processed by the recognizer, while the quadrigram language model was used for the last pass corresponding to a word-lattice rescoring. The word posteriors are computed from these

word-lattices, using acoustic scores from adapted acoustic models and linguistic scores from the 4-gram language model.

B. Training parameters for confidence measures

The parameters that are used to compute the LMBB measure, the AC measure and the MAP measure are estimated from a specific corpus composed of 4h of manually annotated broadcast news from ESTER. This corpus is independent of the training corpus of acoustic and language models and of the training set later refered as MTrain.

C. Corpora used to evaluate the merging techniques

Three corpora are used for the evaluation of the various merging techniques studied in this paper. They come from the official test of the ESTER broadcast news corpus which consists of 10h of continuous speech. Those corpora are noted MTrain, MDev and MTest. MDev is used as a validation set for the MLP and the Sugeno-type fuzzy system. In order to be consistent when defining the test set in an experiment, the choice of the data that are used in this test set (MTest) should be recorded after the data used in the training set. Therefore, MTest is defined by choosing the last 3h30 of the official ESTER test corpus. 3h30 of speech in this database correspond to an approximate amount of 31000 words. It has led to the choice of radio channels that are not contained in MTrain or MDev. As a summary MTrain is composed of 1h of Radio Classique, 1h of France Inter, 1h of France Info and 30 mins of Radio France International (RFI). MDev is composed of 1h of France Culture, 1h of France Info, 1h of France Inter and 30 mins of RFI. MTest is composed of 1h of RFI and 2h30 of Radio Television Marocaine (RTM) and is not a part of MTrain or MDev.

D. Evaluation Metrics

In those experiments, the Confidence Error Rate (CER)[27] is used to assess the relevance of a confidence measure to assess the correctness of a word. The CER is the total number of false accepts and false rejects divided by the total number of hypothesis for a given threshold. It gives an estimation of the classification ability of a confidence measure. The threshold that has led to the minimum CER on MTrain is used when computing the CER on the test set MTest.

Some Detection Error Tradeoff (DET) curves are also shown in order to compare the fuzzy logic techniques used in this article. Those DET curves present the two types of error : the false alarms and the false rejections. In the paper written by Martin [28], the authors show that the DET curve form of presentation is relevant to any detection task where a tradeoff of error types is involved. Thanks to the DET curves, performances of confidence measures in terms of their capabilities to validate correct hypotheses and reject false alarms are evaluated.

E. Single Measures

Table I shows the different measures developed in this paper. The best single measures are the WP measure and the MAP measure which have obtained a CER score of 17.88 on MTest. The mapping technique between the WP measure and the MAP measure preserves the CER score as the piecewise linear transformation does not affect the distribution from the optimal threshold. Those measures alone are relevant to assess the correctness of a word and have a CER which is much lower than the rate of the emitted words that are incorrect (17% on MTrain and 22.3% on MTest). By adding measures that are not affected by pruning heuristics, the aim is then to find the best combination that will lead to a better performance in terms of classification.

TABLE I
RESULTS IN TERMS OF CER FOR THE SINGLE CONFIDENCE MEASURES

Confidence Measure	MTrain (%)	MTest (%)
Incorrect emitted words	17	22.3
WP	14.65	**17.88**
MAP	14.65	**17.88**
AC	16.04	22.18
LMBB	16.05	22.17

F. Results after merging

Several combinations of the four measures were tried. To compare the different techniques, some baselines are computed, which are a simple arithmetic mean, a linear regression between the four single measures, and a MLP.

In order to take advantage of the uncertainty notion from the possibility theory, various fuzzy systems are tried to aggregate the four confidence measures.

Figure 1 illustrates how the Choquet Integral method computes the weights used by the fusion process (section IV). Every path of the lattice leads to a part of the plan where every single measure is associated with a weight p. Depending on which measure is the highest for the current word, it will lead to a different linear interpolation merging the four measures.

For example, let us consider a word w and its four scores coming from each confidence measures. For this word w, we have:

$$m_{WP}(w) < m_{MAP}(w) < m_{AC}(w) < m_{LMBB}(w) \quad (13)$$

The final score $m_{CI}(w)$ that will be associated to that word is:

$$\begin{aligned} & m_{WP}(w) * (p(\{1234\}) - p(\{234\})) \\ & + m_{MAP}(w) * (p(\{234\}) - p(\{34\})) \\ & + m_{AC}(w) * (p(\{34\}) - p(\{4\})) \\ & + m_{LMBB}(w) * p(\{4\}) \end{aligned} \quad (14)$$

Table II summarizes the results of the various combination techniques. The fusions on possibillity measures do not perform well. Every other technique outperforms the arithmetic mean and the linear regression and is more relevant to assess the correctness of a word. As the training corpus is substantial,

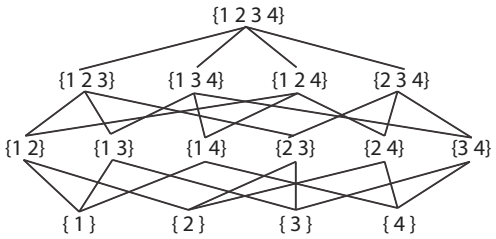

Fig. 1. Lattice for 4 measures. Each path is associated with a weight processed by the Choquet Integral technique on MTrain. $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$ corresponds respectively to the weights associated to the WP, MAP, AC and LMBB measure.

TABLE II
RESULTS IN TERMS OF CER FOR THE VARIOUS MERGING TECHNIQUES.

Aggregated Measure	MTrain (%)	MTest (%)
Arithmetic mean	14.18	17.61
MLP	13.87	**17.44**
Linear Regression m_{LR}	13.92	17.55
Normalised conjonctive fusion	16.04	22.16
Disjonctive fusion	16.04	22.15
Adaptative fusion	16.04	22.17
Sugeno Integral m_{SI}	13.72	**17.44**
Choquet Integral with possibility measures	14.72	21.16
Choquet Integral m_{CI}	13.90	**17.49**

the Choquet Integral does not perform better than the Sugeno Integral but they lead to a good performance in terms of CER.

VI. SELECTION OF THE MEASURES

In order to understand the correlation between the different measures and to better aprehend which measure is important in the aggregation and which one is not, their important index and their interaction index can be good indicators. Those indexes are provided by the Choquet lattice and as the Sugeno and Choquet Integrals are very close in terms of the CER score, they will be used with the Choquet Integral to observe the relevance of selecting measures.

The important index is described in section IV-B3. It describes the correlation between the score given by a measure and the correctness of a word. All the important indexes are computed for each confidence measure and are detailed in table III.

TABLE III
IMPORTANCE INDEX OF THE FOUR MEASURES

Confidence measure	Importance Index
WP	0.517512
MAP	1.930417
AC	0.536041
LMBB	1.016030

The first merging technique simply consists in a linear interpolation of the four measures with the importance indexes used as coefficients. The resulting measure is :

$$m_{II}(w) = \frac{0.517512 m_{WP} + 1.930417 m_{MAP} + 0.536041 m_{AC} + 1.01630 m_{LMBB}}{4} \quad (15)$$

The interaction index described in section IV-B3 is also computed on our data. The results are shown in table IV.

TABLE IV
INTERACTION INDEX OF THE FOUR MEASURES

	WP	MAP	AC	LMBB
WP	0.000000	-0.255104	0.041425	-0.776167
MAP	-0.255104	0.000000	-0.418659	-1.405465
AC	0.041425	-0.418659	0.000000	0.102432
LMBB	-0.776167	-1.405465	0.102432	0.000000

The idea to select the measures that will be kept in the aggregation is to look for the measure associated with the weakest important index and for which the interaction index is lower than the average interaction index.

This led to the results shown in table V.

TABLE V
RESULTS IN TERMS OF CER FOR THE AGGREGATION USING SELECTED MEASURES

Aggregated Measure	MTrain (%)	MTest (%)
Choquet Integral (4 measures)	13.90	**17.49**
Linear regression with important indexes m_{II}	13.82	**17.34**
Choquet Integral (3 measures)	13.98	18.22
Choquet Integral (2 measures)	13.91	**17.20**

Those results show the relevance of the use of important indexes. By using them as coefficients in a simple linear regression, the resulting measure outperforms the results of a MLP and of the two fuzzy systems based on the Choquet and the Sugeno integrals.

Selecting the measures that will be used in the aggregation shows a good performance and leads to the best performance in terms of CER when using only two measures with the Choquet Integral.

The DET curve (see figure 2) confirms that, by selecting only two measures for the Choquet Integral, we obtain a relevant aggregation in terms of its ability to reject false alarms. The aggregation of three measures seems to obtain good results for the validation of correct hypotheses.

VII. DISCUSSION AND CONCLUSIONS

This paper presents how fuzzy logic is relevant in measuring the confidence in a speech processing application.

For applications where the interesting aspect of the confidence measure is its classification ability, the CER improvement is important. For example, one might want a confidence measure that is well able to detect incorrect words in order to increase the amount of training data for the acoustic models in an unsupervised way. According to the CER, the fuzzy logic techniques prove to be very efficient and combine the best of the various individual confidence measures involved on each data subset. Compared to a standard MLP, fuzzy integrals and the Choquet Integral in particular can provide a semantic

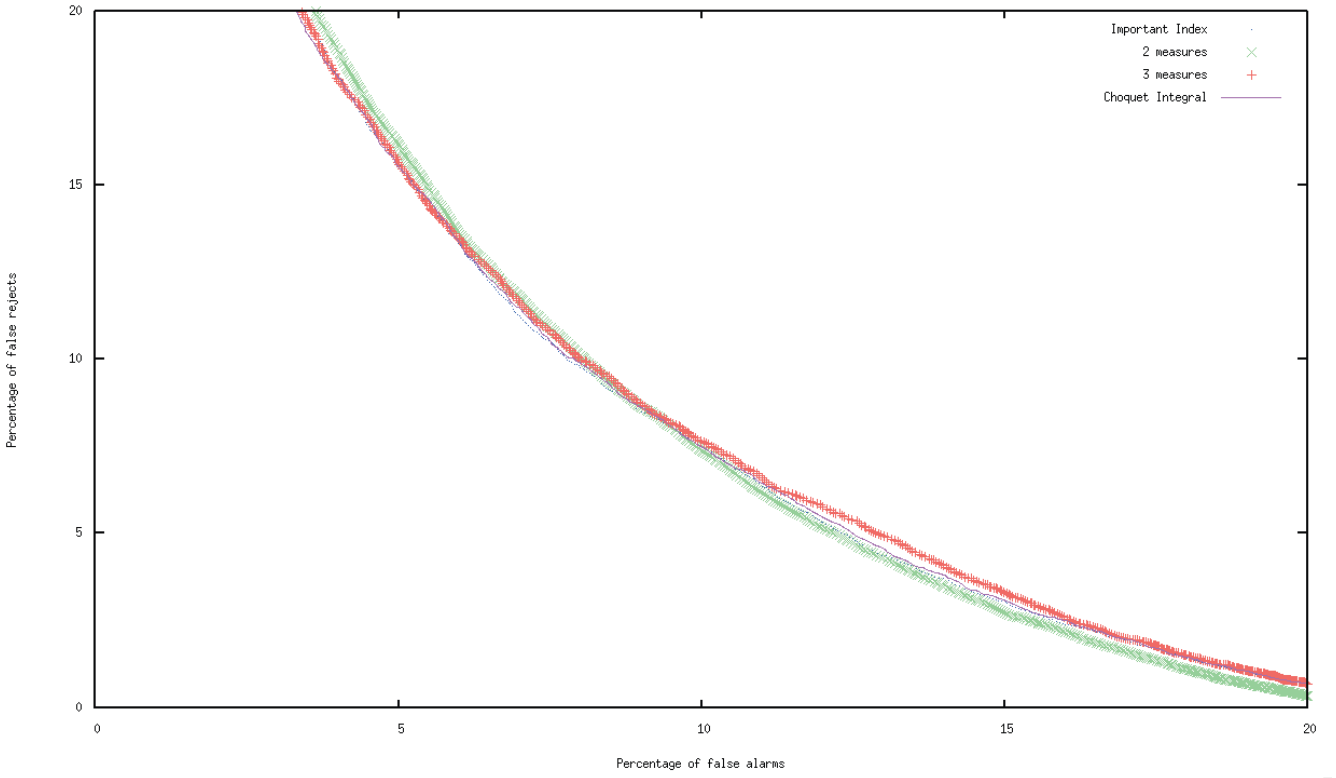

Fig. 2. DET curve for the Choquet Integral with 4, 3 and 2 measures and with the linear regression with important indexes.

interpretation and understanding of the problem. The Choquet integral technique, by selecting only two measures, improves the CER by 4.10% compared to the CER obtained by the best single confidence measure. From DET curves, the selection of the measures has also proven to increase the discriminative ability of confidence measures and to be worthwhile.

In future works, other methods using fuzzy logic will be explored.

REFERENCES

[1] H. Jiang, "Confidence measures for speech recognition: a survey," *Speech Communication Journal*, vol. 45, pp. 455–470, 2005.

[2] D. Charlet, G. Mercier, and D. Jouvet, "On combining confidence measures for improved rejection of incorrect data," in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Aalborg, Danemark, September 2001.

[3] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 23–31, 2005.

[4] R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. Pardo, "Confidence measures for spoken dialogue systems," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May 2001.

[5] F. Metze, T. Kemp, T. Schaaf, T. Schultz, and H. Soltau, "Confidence measure based language identification," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000.

[6] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," in *IEEE Transactions on Systems, Man and Cybernetics*, February 1988, pp. 183–190.

[7] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Aalborg, Danemark, Septembre 2001, pp. 2105–2108.

[8] M. Grabisch and N. J.M., "Classification by fuzzy integral - performance and tests," in *Fuzzy Sets and Systems, Special Issue on Pattern Recognition*, 1994, pp. 255–271.

[9] M. Grabisch, "The application of fuzzy integrals in multicriteria decision making," in *European Journal of Operational Research*, 1996, pp. 445–456.

[10] J.-L. Marichal, "Aggregation of interacting criteria by means of the discrete choquet integral," in *Physica-Verlag GmbH*, 2002, pp. 224–244.

[11] M. Grabisch, "A new algorithm for identifying fuzzy measures and its application to pattern recognition," in *IEEE International Conference on Fuzzy Systems*, 1995, pp. 145–150.

[12] J. Mauclair, Y. Estève, S. Petit-Renaud, and P. Deléglise, "Automatic detection of well recognized words in automatic speech transcriptions," in *LREC, Language Resources and Evaluation*, Genoa, Italy, May 2006.

[13] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Speech Transcription Workshop*, 2000.

[14] G. Evermann and P. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000.

[15] H. Mangu, E. Brill, and S. A., "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[16] M. Laviolette and J. Seaman Jr., "The efficacy of fuzzy representations of uncertainty," *IEEE Transactions on Fuzzy Systems*, vol. 2, no. 1, p. 415, February 1994.

[17] J. Mendel, "Fuzzy logic systems for engineering: a tutorial," *Proceedings of the IEEE*, vol. 83, no. 3, pp. 345–377, March 1995.

[18] D. Dubois, H. Prade, and S. Sandra, "On possibility/probability transformations," in *Proc. of the Fourth IFSA Conference*, Seoul, Korea, 1993, pp. 103–112.

[19] G. Choquet, "Theory of capacities," in *Annales de l'Institut Fourier*, 1953, pp. 131–295.

[20] T. Murofushi and M. Sugeno, “A theory of fuzzy measures : representations, the choquet integral, and null sets,” in *Journal of Mathematical Analysis and Applications*, 1991, pp. 532–549.

[21] J.-S. R. Jang, “ANFIS: Adaptative-network-based fuzzy inference system,” *IEEE Transactions on systems, man and cybernetics*, vol. 23, no. 3, pp. 665–685, May/June 1993.

[22] G. Hernandez-Abrego, G. Hernandez, and J. Marino, “Fuzzy reasoning in confidence evaluation of speech recognition,” *IEEE International Workshop on Inteligent Signal Processing WISP’99*, September 1999.

[23] L. Shapley, “A value for n-person games,” in *Contributions to the Theory of Games, Annals of Mathematics Studies*, 1953, pp. 307–317.

[24] T. Murofushi and S. Soneda, “Techniques for reading fuzzy measures(III): interaction index,” in *Proc. of the 9th Fuzzy System Symposium*, Sapporo, Japan, May 1993, pp. 693–696.

[25] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier, “The ESTER phase II evaluation campaign for the rich transcription of french broadcast news,” in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[26] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, “The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news,” in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[27] F. Wessel, K. Macherey, and R. Schlüter, “Using word probabilities as confidence measures,” in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Seattle, USA, Mai 1998, pp. 225–228.

[28] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,” in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Rhodes, Greece, September 1997, pp. 1895–1898.