# RESCORING UNDER FUZZY MEASURES WITH A MULTILAYER NEURAL NETWORK IN A RULE-BASED SPEECH RECOGNITION SYSTEM

*OPPIZZI Olivier & QUÉLAVOINE Régis*

Laboratoire Informatique d'Avignon
oppizzi@univ-avignon.fr quelavoine@univ-avignon.fr

## ABSTRACT

In this paper, a speech rescoring system is developed on a set of phonetic hypotheses produced by a bottom-up knowledge-based decoder. An original method to automatically compute a fuzzy membership function from top-down acoustic rules statistics is compared with a possibilistic measure. To aggregate the fuzzy degrees into a phonetic score, a mutilayer neural network is trained on the results of all the rules in order to detect how these rules characterize different phonemes and then in order to give a weight to each rule. Rescoring performance of top-down rules for fricatives will be discussed on an isolated-word speech database of French with 1000 utterances pronounced by five speakers.

## 1. INTRODUCTION

Various methodologies for approaching the rescoring problem have been proposed. [2] used a bayesian operator to aggregate probabilist scoring measures and [3] described an HMM rescoring methodology by defining a neural network that takes speech data as inputs and produces as output the probability of a phoneme.

Instead, the goal of the present work is the computation of a fuzzy measure from insufficient statistics in a rule-based isolated-word recognition system, and the use of a neural network operator that takes fuzzy numbers as inputs to assess the significance and the weight of each rule in order to optimize the computation of a fuzzy phonetic score.

## 2. A FUZZY MEMBERSHIP FUNCTION

A bottom-up, rule-based, acoustic-phonetic decoder retrieves and scores the phonetic hypotheses from a speech signal [1]. To improve recognition performances by rescoring the phonetic list, additional knowledge sources (top-down rules) have been determined to verify coarticulation features. Applied to a phonetic hypothesis, each rule returns a numeric parameter related to a fuzzy number via a procedure described in [5] : a fuzzy set of rule parameters is made up and the membership function $C_R()$ is drawn from one-speaker database histograms.

Let $F$ be the class of unvoiced fricatives. In figure 1, a 400-observation histogram $HR1$ is drawn for correct recognition of F-phonemes by a given rule $R$, and a 250-observation histogram $HR2$ is drawn for non-F-phonemes recognized as F-phonemes at bottom-up decoding (the vertical axis is the zero-crossing rate parameters returned by $R$). The few number of observations and the fact that the rule is only applied to a one-speaker database explain why the statistics are not sufficient for generalization, and why a fuzzy membership function is needed.

[4] gives a method to compute a possibilistic function $\pi_i()$ from histogram $HRi$. The aggregation function $\pi() = min(\pi_1, 1 - \pi_2)$ is illustrated in figure 1 : the more is the possibility of an erroneous recognition, the less must be the possibility of a correct recognition.

Our fuzzy function $C_R()$ is more robust to unrelevant histogram variations. In case of null probabilities, the value of ignorance is 0.5. In case of a $HR2$ peak higher than the corresponding $HR1$ peak, the $C_R()$ values are under 0.5. The possibility function $\pi()$ is not able to distinguish between ignorance and a higher $HR2$ peak.

## 3. USE OF A MULTILAYER NEURAL NETWORK TO WEIGHT THE RULES

$c_{ij} = C_{Ri}(P_j)$ are fuzzy numbers which correspond to the degree of certainty to detect the phoneme $P_j$ knowing the result of rule $R_i$. Let $F_j = [c_{ij}]_{0<i<n+1}$ be the fuzzy vector for phoneme $P_j$ where $n$ is the number of rescoring rules. To properly weight fuzzy degrees $c_{ij}$ in order to achieve an optimal phonetic score, we decided to use a three-layer neural network (input, hidden and

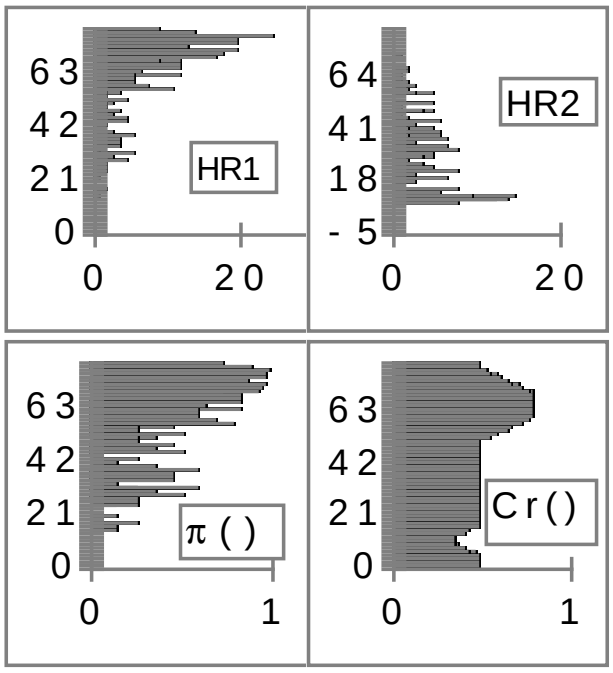


Figure 1: Histograms $HR1$ : good recognitions, $HR2$ : bad recognitions, and the possibilistic function $\pi()$ and fuzzy function $Cr()$, for the rule $R$ designed in order to characterize unvoiced fricatives

output layers) that would recognize a phoneme with $F_j$ the fuzzy vectors as inputs. The analysis of the weights of the network will then give an idea of the significance of the rules.

As an illustration, we tried to separate the fricative phonemes (v, z, j, f, s, ch) from the others. We used a five-speaker corpus composed of examples of 33 french phonemes, described through 36 rules, each rule giving a score in $[0, 100]$. So we have 48944 examples with 6264 fricatives : 828 f, 2159 s, 252 ch, 1665 v, 898 z and 462 j.

### 3.1. Learning procedure

To find which inputs are really relevant, the network must not be too big (adaptation to the corpus and poorer generalization scores), so we pruned all superfluous nodes [8]. Label errors that can be present in the corpus must be detected and overtraining must be avoided, so we used a selective learning algorithm [9].

We could then build a (36, 12, 1) network that learns the whole corpus with a 0.2% error rate (the difference between the expected and the network response is over 0.5 in $[0, 1]$). These errors are coherent with usual results for the fricative recognition problem : 5 d, 13 t, 2 p (these are occlusives that may have fricative features depending on the context) and 6 j, 6 z, 20 f, 25s, 29 v. The v is actually the most difficult fricative phoneme to recognize because of its shortness (its pronounciation is greatly affected by neighbouring vowels).

### 3.2. Input hierarchy

We know since [7] that multilayered neural networks in classification are making a discriminant analysis of the inputs, so somewhere in their structure is the information we are looking for. In [6], we can measure the features significance by testing the effect of a variation of each input for each example, that means a lot of calculus. We reach the same results (0.92 correlation) with a much more efficient method inspired by [10] : $\frac{\partial output}{\partial input}$.

Since each input varies in $[0, 100]$, the only weights are significant of the inputs importance, and their signs give an idea of the activation / inhibition power. Using the sigmoïd $f(x) = \frac{1}{1+e^{-x}}$ as the activation function, we have $0 \leq f'(x) \leq 0.25$, so :

$$\sum_{l \in L_j} (V_l \cdot W_{jl}) \leq 16 \cdot \frac{\partial output}{\partial input_j} \leq \sum_{l \in K_j} (V_l \cdot W_{jl})$$

with $W_{ij}$ the weight from input $i$ to hidden node $j$, $V_i$ the weight from hidden node $i$ to the output, and $L_j = \{l \ / \ V_l \cdot W_{jl} < 0\}$, $K_j = \{l \ / \ V_l \cdot W_{jl} > 0\}$.

We know then the activation ($K_j$) / inhibition ($L_j$) power for each input and their range give the discrimination power ($K_j + |L_j|$).

### 3.3. Relevance of the rules

For our fricative recognition problem, we can quantify which rules are the most efficient, and which ones are superfluous. This quantification allows us to establish the weight of each fuzzy number as input when aggregating $F_j = [c_{ij}]$ to a score for phoneme $j$:

- rules 1, 3, 4 and 21 each represent less than 0.5% of the network weights (instead of an average 3%). Indeed, these rules were designed to verify non-fricative features. Surprisingly, one rule which was not initially proposed to distinguish fricatives reaches 11%. Thus, results of the network can help the phoneticians to find other regularities in the speech signal. Occlusive rules 30-24 give good results as well to distinguish the fricatives. Conversely, the zero-crossing rate fricative

rule gives weak results (1%), showing that it can not be trusted in a multi-speaker environment (too wide variations).

- fricatives rules 36, 20 and rule 9 activate a fricative response, whereas occlusive rules 30-24 and other rules 19-16 are just used for inhibition purposes. The remaining rules give both effects, and so do not accurately characterize fricatives.

After pruning the weakest rules, we kept a very low error rate (121 errors instead of 106 out of 48944 examples). Althought working with the whole corpus diminishes the risk of biased learning results, a perspective is to repeat our experiments by learning on a limited corpus and by testing the remaining data to study the effect of inputs selection upon generalization. But our goal here was just to find how to weight the rules, and not to build a neural network based recognition system.

## 4. RESCORING PERFORMANCE

The isolated-word recognition corpus consisted of 1000 words selected from the BDLEX database and pronounced by five speakers (these examples are not in the corpus used for the neural network training). To reduce a phonetic list and then to diminish lexical hypotheses, the experiment consisted of testing the rescoring ability of 36 top-down rules on fricatives generated by a bottom-up decoder. This one produced 3262 correct fricatives and 3041 hypothesized fricatives which are not fricatives actually.

Table 1 shows the phonetic rejection rates without rescoring (from bottom-up scores), table 2 gives the rejection rates with experimental weights $|c_{ij} - 0.5|$ (named OWA rescoring) and table 3 the rejection rates when aggregating the fuzzy degrees $c_{ij}$ with neural weights. The right column of the tables shows the rejection rates of erroneous fricative hypotheses (we expect these rates to be high in a robust rescoring system), and the middle column shows the rejection rates of correct fricative hypotheses (we expect these rates to be low).

| Thresh. | fric. reject. | non fric. reject. |
|---|---|---|
| 40 | 0% | 3.2% |
| 61 | 9.5% | 26.4% |
| 71 | 27% | 50.4% |

Table 1: Rejection rates on fricatives without rescoring

The first table shows bad results if no rescoring procedure is used. With neural rescoring, the phonetic list can be reduced down to 25% of its wrong fricatives

| Thresh. | fric. reject. | non fric. reject. |
|---|---|---|
| 22 | 0% | 1% |
| 28 | 2.5% | 25% |
| 35 | 6.7% | 52% |

Table 2: Rejection rates on fricatives with OWA rescoring

| Thresh. | fric. reject. | non fric. reject. |
|---|---|---|
| 36 | 0% | 2.8% |
| 40 | 1.8% | 25.1% |
| 71 | 5.9% | 51.8% |

Table 3: Rejection rates on fricatives with neural rescoring

(column "non fric. reject.") with less than 2% correct phoneme rejection (column "fric. reject."). It proves that additional acoustic-phonetic rules provide a set of discriminant information to recognize the fricatives.

Besides, the neural rescoring procedure was observed to give a slight but significant improvement (with a 95% confidence interval, $\Delta = \pm 0.8$) over the OWA rejection method (more correct fricatives are rejected by this latter method). Note that the neural weighted sum provided better performance if average fuzzy degrees ($c_{ij} = 0.5$) are ignored (they mean either ignorance or high uncertainty).

Nevertheless, the multi-speakers OWA rejection rates are encouraging since this method needs only a small one-speaker corpus to be effective. The fuzzy approach allows a robust modelling stage of an incomplete body of certainty.

It is also interesting to note that the so-called fuzzy acoustic-phonetic decoder applies an automatic least-commitment decision-making strategy once the acoustic rules are determined. Expert decision thresholds and meta-rules are avoided in such a system.

## 5. CONCLUSION

An original method to automatically compute a fuzzy membership function from insufficient statistics (a limited corpus of one-speaker speech signals) has been proposed in the attempt to assess phonetic rescoring performance in a rule-based speech recognition system. A weighted sum operator was chosen to aggregate fuzzy numbers into a phonetic score. The use of a neural network was proposed to establish the weights of the rules when computing the aggregation operator.

The neural network allows to distinguish relevant and unrelevant rules for a given phonetic class, that is, rules which give a high discriminative rate in a multi-speaker environment. Experimental results on fricatives demonstrated that the neural network provides a way to perform an encouraging fricative rejection rate.

The future of this work will lead us to verify these results on other phonetic classes and to apply the fuzzy decoding approach to a "state-of-the-art" hybrid recognition system with a stochastic agent as the bottom-up decoder and with neural networks as additional top-down agents.

## 6. REFERENCES


[1] Gilles P. : *Décodage phonétique de la parole et adaptation au locuteur* . Thèse de l'Université d'Avignon, 1993.

[2] Rivlin Z. : *A Confidence Measure for Acoustic Likelihood Scores.* EuroSpeech'95, vol.1, pp.523-526, Madrid, 1995.

[3] Zavaliagkos G., Zhao Y., Schwartz R. & Makhoul J. : *A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition.* IEEE Transactions on Speech and Audio Processing, vol.2(1), pp.151-159, 1994.

[4] Dubois D., & Prade H. : *Théorie des possibilités : application à la représentation incertaine des connaissances dans les systèmes intelligents.* MASSON (ed), Paris, 1988.

[5] Oppizzi O., Fournier D., Gilles P., & Méloni H. : *A Fuzzy Acoustic-Phonetic Decoder for Speech Recognition.* ICSLP, Philadelphia, 1996.

[6] Fechner T. & Hinze A. : *Delta Analysis : a method for the determination of input feature significances in neural networks.* NeuroNîmes'93, pp.393-398, Nîmes, 1993.

[7] Gallinari P., Thiria S. & Fogelman Soulié F. : *Multilayer perceptrons and data analysis.* IEEE 2nd ICNN, vol.1, pp.391-401, San Diego, 1988.

[8] Nocera P. & Quélavoine R. : *Diminishing the number of nodes in multilayered neural networks.* IJCNN'94, vol.7, pp.4421-4424, 28jun-2jul94, Orlando, 1994.

[9] Quélavoine R., Nocera P. & Di Martino M. : *Multilayered neural networks and errors in learning corpus.* INNS WCNN'96, pp.287-290, 20-25sep96, San Diego, 1996.

[10] Yoda M., Baba K. & Enbutu I. : *Explicit representation of knowledge acquired from plant historical data using neural networks.* IJCNN'91, vol.3, pp.155-160, San Diego, 1991.