COMBINATION OF ACOUSTIC CLASSIFIERS BASED ON DEMPSTER-SHAFER THEORY OF EVIDENCE

Fabio Valente and Hynek Hermansky

IDIAP Research Institute, CH-1920 Martigny, Switzerland
Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland
{*fabio.valente,hynek.hermansky*}*@idiap.ch*

ABSTRACT

In this paper we investigate combination of neural net based classifiers using Dempster-Shafer Theory of Evidence. Under some assumptions, combination rule resembles a product of errors rule observed in human speech perception. Different combination are tested in ASR experiments both in matched and mismatched conditions and compared with more conventional probability combination rules. Proposed techniques are particularly effective in mismatched conditions.

Index Terms— Dempster-Shafer theory, Classifier combination, Multi-Stream ASR.

1. INTRODUCTION

Multi-stream speech recognition approaches where individual information streams are formed by using evidence from different elements of the signal are becoming a norm in the ASR community (e.g. multi-band [1],[2], feature combinations [3]). In this paper, we study combinations of posterior probabilities of phonemes derived from different input speech representations. The probabilities are estimated by a multi-layer perceptron (MLP) trained on phoneme-labeled data.

In literature, many papers have already addressed the problem (e.g. [4]) considering combination rules like sum, product, maximum and minimum rules. Anyway combination has always been considered in the framework of classical probability theory. We study here a combination rule based on Dempster-Shafer theory of evidence ([5]) which can be considered an extension of Bayesian probability. Main advantage of this framework is the explicit representation of ignorance. DS theory has already been investigated in speech recognition (e.g. [6]) but this is probably the first attempt to use it for combination of information coming from different acoustic streams. Furthermore, under some assumption, DS combination rule is similar to what is known in the speech recognition community as the Fletcher's "product of errors" (see [7],[8]).

The paper is organized as follows: section 2 gives some generalities on Theory of Evidence, section 3 draws a parallel between DS combination rule and product of errors, section 4 describes how to transform output of an MLP into a BPA, section 5 describes how to combine BPAs coming from different MLP, and finally section 6 describes experimental results.

2. THE DEMPSTER-SHAFER THEORY OF EVIDENCE

The Dempster-Shafer (DS) Theory of Evidence (see [5]) allows representation and combination of different measures of evidence. It can be considered as a generalization of the Bayesian framework and permits the characterization of uncertainty and ignorance.

Let $\Theta = \{\theta_1, ..., \theta_k\}$ be a finite set of mutually exclusive and exhaustive hypotheses refereed as singletons. Θ is referred as *frame of discernment*. Let 2^Θ be the power set of Θ i.e. the set of all subsets of Θ. A *basic probability assignment* (BPA) is a function m from 2^Θ to $[0, 1]$ such that

$$m : 2^\Theta \rightarrow [0, 1], \quad \sum_{A \subset \Theta} m(A) = 1 \text{ and } m(\oslash) = 0 \tag{1}$$

$m(A)$ can be interpreted as the amount of belief that is assigned exactly to A and not to any of its subsets. In probability theory, a measure is assigned only to atomic hypothesis $m(\theta_i)$ while in DS Theory it can be assigned to a set A without any further commitment on the on the atomic hypothesis that compose A. The situation of total ignorance is represented by $m(\Theta) = 1$. On the other hand, if we set $m(\theta_i) \neq 0$ for all θ_i and $m(A) = 0$ for all $A \neq \theta_i$, we recover the probability theory.

Let $\neg A$ be complementary set of A i.e. the set $\{\Theta - A\}$. In DS Theory, $m(A) + m(\neg A) < 1$ (contrarily to probability theory), which means that we can consider an amount of belief that is not attributed to an hypothesis nor to its negation. In other words, "we don't need to over-commit when we are ignorant".

The function that assigns to each subset A, the sum of all basic probability numbers of its subset is called *belief function* or *credibility* of A:

$$Bel(A) = \sum_{B \subset A} m(B) \tag{2}$$

Subset A for which $m(A) > 0$ are called *focal elements* and their union is called *core*. A belief function is defined as *vacuous* if it has only Θ as focal element. A belief function is defined as *simple support* function if it has only one focal element in addition to Θ and Bayesian if its focal elements are singleton.

In an analogous way, *Plausibility* of an hypothesis A is defined as:

$$Pl(A) = 1 - Bel(\bar{A}) = \sum_{B \cap A \neq 0} m(B) \tag{3}$$

and it measures to what extent we fail to doubt in A. Another interesting point in DS Theory is how two different belief functions Bel_1 and Bel_2 over the same frame of discernment are combined into a single belief function. Dempster's rule states that Bel_1 and Bel_2 must be combinable i.e. their cores must not be disjoint. Given m_1 and m_2 BPAs associated with Bel_1 and Bel_2 this condition can be

expressed as $\sum_{A\cap B=\oslash} m_1(A)m_2B < 1$. In this case m_1 and m_2 can be combined as:

$$m(\oslash) = 0, \;\; m(\theta) = \frac{\sum_{A\cap B=\theta} m_1(A)m_2(B)}{1-\sum_{A\cap B=\oslash} m_1(A)m_2(B)} \tag{4}$$

and $m(\theta)$ is a BPA. The belief function given by m is called orthogonal sum of Bel_1 and Bel_2 denoted as $Bel_1 \oplus Bel_2$ (m as well is denoted as $m_1 \oplus m_2$). DS orthogonal sum is both associative and commutative. Given two belief functions Bel_1 and Bel_2, if Bel_1 is vacuous, then $Bel_1 \oplus Bel_2 = Bel_2$; if Bel_1 is Bayesian, then $Bel_1 \oplus Bel_2$ is also Bayesian.

Let us consider now the case of orthogonal sum between two simple support belief functions Bel_1 and Bel_2 with focus $A \neq \Theta$ i.e. $m_1(A) = s_1, m_1(\Theta) = 1-s_1, m_2(A) = s_2, m_2(\Theta) = 1-s_2$. Applying DS orthogonal sum (4), we obtain:

$$m(\Theta) = (1-s_1)(1-s_2), \;\; m(A) = 1-(1-s_1)(1-s_2) \tag{5}$$

In words, in case of simple support belief functions, the total ignorance is the product of ignorances of single belief. In next section, we draw a parallel with product of errors.

3. PRODUCT OF ERRORS

Work of Fletcher ([7]) on human processing of speech suggests that humans process speech in different frequency sub-bands independently. Combination of processing from each sub-band is done in such a way that total error is equal to product of errors in different sub-bands. In other words, to recognize correctly a phoneme it is enough to recognize it correctly in one of the available sub-bands.

Those findings suggested as possible combination rule of classifiers based on different acoustic evidence, the product of errors (PoE). Let us denote with p_1 and p_2 the probability of correct recognition of a phoneme for two different acoustic streams, according to PoE, the combined probability of those classifiers should be $p = 1-(1-p_1)(1-p_2)$. It is evident the analogy in between previous expression and results from expression (5) with the difference that in theory of evidence we should talk about "product of ignorances" rather then "product of errors". Anyway, as we will verify in the experimental section, combination according to PoE does not provide results comparable to classical classifiers combination rules; on the other hand, "product of ignorances" gives good results compared to other rules.

4. FROM MLP OUTPUT TO BASIC PROBABILITY ASSIGNMENT

DS theory represents an interesting alternative to classical probability framework for combining different classifiers and it has already been largely studied in the machine learning community (e.g. see [9]). Main weakness of DS theory is the fact that results are strongly sensitive on the choice of the Basic Probability Function. Thus DS combination rule has a certain degree of heuristic depending on the type of classifier we aims at combining.

We will focus on combination of outputs from different Neural Networks. In [10] and [11], multiple neural nets outputs are combined using DS orthogonal sum for handwriting recognition applications. The main question is how to choose an effective BPA. Each output from the a neural net is considered as a source of information (a belief) that induces a frame of discernment. If we denote with θ_i the $i-th$ output of the MLP, focal elements of the corresponding BPA will be $m_i(\theta_i)$ i.e. the belief we have in the hypothesis associated with the $i-th$ output, $m_i(\neg\theta_i)$ i.e. the belief we have in the complementary of this hypothesis and $m_i(\Theta)$ i.e. the ignorance associated with this hypothesis. In [10], BPA are estimated respectively according to recognition rate, error rate and rejection rate of each Neural Net output while in [11], they are estimated according to different kind of distances between MLP outputs and some reference vectors.

We consider the output of a Neural Network trained in order to estimate posterior distributions for a target class (i.e. a phoneme posterior) [12]. Let us consider a phoneme set $\Theta = \{\theta_1, ..., \theta_k\}$ and a trained Neural Net that produces target posteriors $\{p_1 = p(\theta_1|X), ..., p_k = p(\theta_k|X)\}$ with $\sum_i p_i = 1$ where X is an observation vector. First problem we have to deal with is how to transform the probabilistic output of the MLP into a BPA. With DS formalism, the probabilistic output can be represented by the following BPA $m(\theta_i) = p_i \;\; \forall i$ and $m(\Theta) = 0$ i.e. all belief is attributed to atomic hypotheses (phonemes) and no belief to the ignorance. To quantify the degree of ignorance of the MLP output, a natural choice is the use of the entropy of the output $H = \sum_i^k p_i log(p_i)$. Ignorance is supposed to be total (i.e. $m(\Theta) = 1$) when entropy of the output achieves its maximum value $H_{max} = \sum_i^k \frac{1}{k} log(\frac{1}{k})$. Under those considerations a possible choice for a BPA is represented by:

$$m_i(\theta_i) \;\; = \;\; \alpha\, p_i \qquad m_i(\Theta) = 1-\alpha\, p_i = 1-m_i(\theta_i) \tag{6}$$

$$\text{with } \alpha \;\; = \;\; (1-\frac{H}{H_{max}})^{\gamma} \tag{7}$$

When the entropy H is zero, ignorance $m_i(\Theta)$ is equal to $1-p_i$ while when entropy is maximum ignorance $m_i(\Theta) = 1$. Choice of function (7) is heuristic; exponent factor γ is supposed to better fit ignorance estimation to entropy measure because ignorance should may not be a linear function of the entropy. BPAs as defined in (6) are simple support functions and we refer to them as BPA1.

Anyway other BPAs can be defined in which we further add information on the complementary set $\neg\theta_i$. For instance we could define a new BPA as:

$$m_i(\theta_i) = \alpha\, p_i \qquad m_i(\neg\theta_i) = \alpha(\sum_{j\neq i} p_j) \tag{8}$$

$$m_i(\Theta) = 1-m_i(\theta_i)-m_i(\neg\theta_i) \tag{9}$$

In this case each MLP output is supposed to provide information on both phoneme i and set of phonemes $\Theta - i$. Contrarily to probability theory, they do not sum to one because a certain amount of belief is supposed to be assigned to all phoneme set Θ. We refer to BPAs (8-9) as BPA2.

Finally a third set of BPA can be directly derived from orthogonal sum of BPAs (6). In fact BPA from each MLP output as defined in (6) are combinable; applying orthogonal sum (4) ($\oplus_i m_i$) a new set of BPA can be directly obtained:

$$m(\theta_i) = m_i(\theta_i)\prod_{j\neq i}(1-m_j(\theta_j))/Z \tag{10}$$

$$m(\neg\theta_i) = (1-m_i(\theta_i))\prod_{j\neq i}(1-m_j(\theta_j))/Z \tag{11}$$

$$m(\Theta) = \prod_{j}(1-m_j(\theta_j))/Z \tag{12}$$

$$Z = 1-m_i(\theta_i)(1-\prod_{i\neq j}(1-m_j(\theta_j)) \tag{13}$$

We refer to set of BPAs (10-13) as BPA3.

In this section, we described three different ways of associating a basic probability assignment on a frame of discernment induced by a MLP output. In next section we describe how to combine two different BPAs obtained trough two different Neural Networks.

5. DS THEORY FOR CLASSIFIERS COMBINATION

Let us consider now the case in which we have two different Neural Networks and their corresponding BPA obtained in one of the three ways described in previous section. Those BPA can now be combined applying orthogonal sum (4). In case of simple support functions (i.e. BPA1), we must combine BPA with only one focal element. Given two MLP a and b and correspondent BPA $m_a(\theta_i) = s_a, m_a(\Theta) = 1 - s_a, m_b(\theta_i) = s_b, m_b(\Theta) = 1 - s_b$, orthogonal sum $m_a \oplus m_b$ gives:

$$\begin{aligned} m(\Theta) &= m_a(\Theta)m_b(\Theta) = (1 - s_a)(1 - s_b) && (14) \\ m(\theta_i) &= m_a(\theta_i)m_b(\theta_i) + m_a(\theta_i)m_a(\Theta) + m_b(\theta_i)m_a(\Theta) \\ &= 1 - (1 - s_a)(1 - s_b) && (15) \end{aligned}$$

Similarity of expressions (14 - 15) with product of errors rule are quite obvious with the difference that in this case combination rule consider product of "ignorance" instead of errors.

In case of BPA2 and BPA3, combination rule must handle as well the set $m(\neg\theta_i)$; orthogonal sum gives:

$$\begin{aligned} m(\theta_i) &= \{m_a(\theta_i)m_b(\theta_i) + m_a(\theta_i)m_a(\Theta) + \\ &+ m_b(\theta_i)m_a(\Theta)\}/Z && (16) \\ m(\neg\theta_i) &= \{m_a(\Theta)m_b(\neg\theta_i) + m_b(\Theta)m_a(\neg\theta_i)\}/Z && (17) \\ m(\Theta) &= \{m_a(\Theta)m_b(\Theta)\}/Z && (18) \\ Z &= 1 - m_a(\neg\theta_i)m_b(\theta_i) - m_b(\neg\theta_i)m_a(\theta_i) && (19) \end{aligned}$$

Combination rules (14 - 15) and (16 - 19) show how to combine BPA from two different MLP into a single BPA. Those rules can be easily extended to more then two classifiers because they are associative.

6. EXPERIMENTS

We investigate the use of DS theory of evidence for combining output of Neural Networks in data-driven feature extractions for ASR. Results are compared with classical combination rules like product and sum.

Data driven feature extraction methods aims at estimating directly from data, features that are used in the recognition process. An effective and well established technique consists in estimating phoneme posteriors using a Neural Network (see [13]). Phoneme posteriors are further processed trough a logarithmic function and a Karunen-Loeve Transform (KLT) before using them as features in the classical HMM/ GMM framework.

Database we used for recognition experiments is the *OGI-Numbers 95* while MLP is trained using 3 hours of hand-labeled speech from the *OGI-Stories* database. Phoneme set is constituted by 29 English phonemes. Two different posterior streams are considered: TANDEM-PLP posterior ([13]) and Multi-RASTA posterior ([14]).

In case of TANDEM-PLP posteriors, MLP input is a vector of 9 consecutive frames of PLP features. In case of Multi-RASTA posteriors, MLP input is a segment of one second critical band energies filtered through a set of multi resolution filters. Those two streams are supposed to capture short and long term dependencies in two different features set. We will consider combination of those different streams according to DS theory.

Multi-RASTA features are inherently robust to linear distortion of the signal [14]. On the other side, Tandem-PLP features are seriously affected by this distortion. To verify the effectiveness of the combination techniques, we study performances of combination when a first order preemphasis filter with $\alpha = 0.95$ is applied to the test data set.

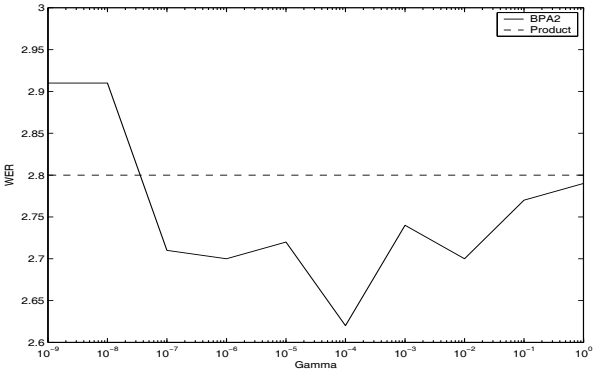

Fig. 1. Performance of combination rule BPA2 function of the factor γ in matched conditions.

Table 1 reports TANDEM-PLP and Multi-RASTA performances in terms of WER in case of matched and mismatched conditions. While Multi-RASTA features hold the performance even in mismatched conditions, TANDEM-PLP are seriously affected.

	Matched	Mismatched
TANDEM-PLP	3.7%	9.7%
Multi-RASTA	3.5%	3.5%

Table 1. WER for TANDEM-PLP and Multi-RASTA features in matched and mismatched conditions.

In the following, we study different combination rules for the two posterior stream. Combined posterior are converted into features using a logarithmic transform and then a KLT transform. Classical way of combining posterior are the sum rule and the product rule (e.g. [4],[15]). We also consider the product of errors rule, directly applied on posterior estimation and inverse entropy weighting (IEW)[16]. In addition to those, we consider combination trough DS theory. When DS theory of evidence is applied, posterior distributions are first transformed into BPA using rules BPA1, BPA2 and BPA3 as described in section 4. BPA from different posterior streams are then combined together using rules described in section 5: BPA1 is combined using rules (14 - 15) (for simple support functions) while BPA2 and BPA3 are combined using rules (16 - 19).

Table 2 shows Word Error Rates for different combination techniques in matched and mismatched conditions. In clean conditions combination of posteriors gives always better results than each posterior stream independently.

Out of the combination rules based on traditional probability theory, product holds the best performance, while product of errors gives the higher error rate. In mismatched conditions, product rule gives same performance of the best feature stream, while sum and product of errors give inferior results.

Let us now consider results from DS combination rules. Out of the three proposed combination framework, the best performing BPA2 is giving 7% improvement in matched conditions and 9% improvement in mismatched conditions w.r.t. product rule.

	Sum	Prod	PoE	IEW	BPA1	BPA2	BPA3
Matched	3%	2.8%	3.1 %	2.9%	2.8%	2.6 %	2.8%
Mismatched	4.1%	3.5%	4.5%	3.8%	3.5%	3.2%	3.5%

Table 2. WER for different combination rules in matched and mismatched conditions. Sum, Prod (product),PoE (product of errors), IEW (inverse entropy weighting).

BPA1 and BPA3 performances are similar to those obtained using product rule. Combination rules BPA1 and BPA3 give very similar results indicating that merging evidence from different outputs of the same MLP does not give any improvement in our experiments.

Many other approaches for combining MLP outputs according to entropy measures have been considered in the past (e.g. [16],[15]). Combination rules still are product rule or sum rule but they are weighted according to some functions of the entropy. In our approach entropy is used to determine the amount of belief from a given MLP output that must be discarded i.e. assigned to the ignorance hypothesis. DS orthogonal sum 4 in the general case cannot be re-conducted into any of those rules.

The most questionable part is the way we transform the output of a probabilistic classifier (i.e. a MLP) into Basic Probability Assignment. Our choices are somehow heuristic and must be further investigated. The use of the entropy is a natural way of representing ignorance but there is no reason for supposing that ignorance should be a linear function of the entropy. As solution to this problem, we choose the function (7) with a correction factor γ. This factor has actually an impact on the final performance of the combination. Figure 1 plots WER in matched conditions as a function of γ for BPA2. WER are sensitive to the value of γ even if there are some intervals in which DS combination performs consistently better than sum or product rules.

7. CONCLUSIONS

In this paper, we present a method for combining output from different neural networks based on Dempster-Shafer Theory of Evidence. Main appeal of this theory is the possibility of representing ignorance. Under certain assumptions (see section 4), DS combination rule show analogies with what was found by Fletcher in his speech perception experiments.

Three different rules for transforming MLP outputs into belief are presented. DS combination rule is tested in recognition experiments and compared with classical combination rules (sum, product and product of errors) both in matched and mismatched conditions. In matched conditions, all combination rules outperforms individual feature streams. Best combination rule is BPA2 while PoE is the worst one. On the other side, product of "ignorances" (i.e. BPA1) shows similar results as the product rule. In mismatched conditions, we would like to have at least a performance equal to the performance of the best feature stream. In case of product rule and BPA1 this is achieved. BPA2 is still achieving error rate lower than the one provided by the best feature stream meaning that it is able to extract useful informations from both streams. Sum and PoE rules are giving error rate higher than those achieved by the best feature stream.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 and by the EU under the grant DIRAC IST 027787. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

9. REFERENCES

[1] Bourlard H. and Dupont S., "A new asr approach based on independent processing and re-combination of partial frequency bands.," *Proc. ICSLP 96*.

[2] Hermansky H., Tibrewala S., and Pavel M., "Towards asr on partially corrupted speech," *Proc. ICSLP 1996*.

[3] Janin A., Ellis D., and Morgan N., "Multi-stream speech recognition: Ready for prime time," *Proc Eurospeech-1999*.

[4] Kirchhoff K. and Bilmes J., "Combination and joint training of acoustic classifiers for speech recognition.," in *ISCA ITRW Conference on Automatic Speech Recognition, Paris*, 2000.

[5] Shafer G., *A mathematical theory of evidence.*, Princeton, MIT Press, 1976.

[6] Kobayashi T., "An application of dempster and shafer's probability theory to speech recognition," *The Journal of the Acoustical Society of America.*, vol. 100 (4), October 1996.

[7] Fletcher H., *Speech and Hearing in Communication.*, Krieger, Hew York, 1953.

[8] Allen J.B., *Articulation and Intelligibility*, Morgan and Claypool, 2005.

[9] Mandler E.J. and Schurman J., "Combining the classification results of independent classifiers based on dempster/shafer theory of evidence.," *Pattern Recognition and Artificial Intelligence*, vol. X, pp. 381–393, 1988.

[10] Xu L., Kryzak A., and Suen C.Y., "Methods of combining multiple classifiers and their applications to handwriting recognition.," *IEEE transactions on Systems, Man and Cybernetics*, vol. 22(3), pp. 418–435, 1992.

[11] Galina L. R., "Combining the results of several neural network classifiers.," *Neural Networks*, vol. 7(5), pp. 777–781, 1994.

[12] Bourlard H. and Morgan N., *Connectionist Speech Recognition - A Hybrid Approach.*, Kluwer Academic Publishers, 1994.

[13] Hermansky H., Ellis D., and Sharma S., "Connectionist feature extraction for conventional hmm systems.," *Proceedings of ICASSP*, 2000.

[14] Hermansky H. and Fousek P., "Multi-resolution rasta filtering for tandem-based asr.," in *Proceedings of Interspeech 2005*, 2005.

[15] Hagen A., *Robust speech recognition based on multi-stream processing*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, December 2001.

[16] Misra H., Bourlard H., and Tyagi V., "Entropy-based multi-stream combination," in *Proceedings of ICASSP*, 2003.