IDIAP RESEARCH REPORT

Multi-stream Features Combination based on Dempster-Shafer Rule for LVCSR System

Fabio Valente [a], Jithendra Vepa [a], Hynek Hermansky [a]

IDIAP–RR 06-61

March 2007

published in
Interspeech 2007

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-61

Multi-stream Features Combination based on Dempster-Shafer Rule for LVCSR System

Fabio Valente, Jithendra Vepa, Hynek Hermansky

March 2007

published in
Interspeech 2007

Abstract. This paper investigates the combination of two streams of acoustic features. Extending our previous work on small vocabulary task, we show that combination based on Dempster-Shafer rule outperforms several classical rules like sum, product and inverse entropy weighting even in LVCSR systems. We analyze results in terms of Frame Error Rate and Cross Entropy measures. Experimental framework uses meeting transcription task and results are provided on RT05 evaluation data. Results are consistent with what has been previously observed on smaller databases.

1 Introduction

Multi-stream speech recognition uses fusion of information coming from different elements of the signal. In the framework of multi-stream ASR, several methods have been proposed e.g. multi-band processing [1],[2] and feature combination [3]. A popular approach to multi-stream system is combining results from several classifiers trained separately on different feature streams or on different frequency band. This combination is generally done according to some probabilistic rules, possibly also taking into account confidence measures associated with individual feature streams ([4]).

In our previous work [5], we showed that efficient way of combining information from different streams is Dempster-Shafer rule [6]. This method is based on theory of evidence which is a generalization of probability theory and allows explicit representation of ignorance. Dempster-Shafer rule has been largely investigated in the pattern recognition community e.g. [7],[8] but application to speech recognition and particularly to LVCSR system has not yet been considered. In [5], experiments were run on a small vocabulary task; here we extend our previous work on LVCSR system for meeting transcription.

Results on RT05 [9] evaluation data shows that out of the considered rules, DS combination still hold the best performance.

The paper is organized as follows: in section 2 we describe data driven feature extraction, TANDEM and MRASTA features, section 3 introduces basic concepts of theory of evidence, section 4 describes how we combine output from two Neural Networks according to Dempster-Shafer combination rule, section 5 describes results of different kind of combination in a meeting transcription task on RT05 evaluation data and finally we present conclusion for this work.

2 Neural Network Feature Extraction

In this section we briefly discuss the use of NNs for performing data driven feature extraction.

Neural Networks are very common tools for classification problems. In a multi-class problem, NN can be trained such that the output approximates class posterior probabilities (see [10]).

In speech recognition, targets are represented by phonetic units, thus NN estimates posterior probability of a given phoneme set. A speech segment can be turned in a *posteriogram* i.e. a representation of the posterior probability of phonemes for each time frame. Ideally a well trained NN will activate an output unit when a given spectro-temporal pattern is presented as input.

The use of 9 consecutive frames of PLP features as input to the net, was first proposed in [11] in the context of hybrid system. In order to use Neural Network output in HMM systems, TANDEM approach was proposed in [12]: posterior probability are gaussianized using logarithm and then decorrelated using a KLT transform. Such features can be used in HMM system alone or in concatenation with classical spectral features. A different approach using longer time segments was proposed in [13] and referred as MRASTA features. In this case, critical band energies are pre-processed using a Multi-resolution set of zero-mean filters . The pre-processed time trajectories are then used as input to a Neural Network. MRASTA features consider quite long temporal context compared to 9 frames PLP and are very robust to linear distortion and noise.

As preliminary experiment, we trained both 9 frames PLP and MRASTA Neural Networks on 110hrs of meetings data (for details on the data set see section 5). Data are phonetically labeled trough forced alignment using the LVCSR system described in [14]. Frame error rate for both NN shows very similar values: 34.6 % (9 frames PLP) 36.3% (MRASTA). On the other side analysis of confusion matrices show that errors from the two classifiers are very different. Figure 1 plots relative difference of confusion matrix diagonals : out of the 46 phonetic targets, in two cases 9 frames PLP and MRASTA have similar performances, in 11 cases 9 frames PLP outperforms MRASTA and in the remaining phonemes MRASTA outperforms 9 frames PLP. It is thus worthy considering combination of Neural Network outputs.

Posterior probabilities from Neural Networks can be combined according to classical probability rules e.g. sum, product, or inverse entropy. The novelty of this paper consists in investigating a non-probabilistic combination rule based on Dempster-Shafer theory of evidence on LVCSR system.

In next section, we introduce basics of theory of evidence.

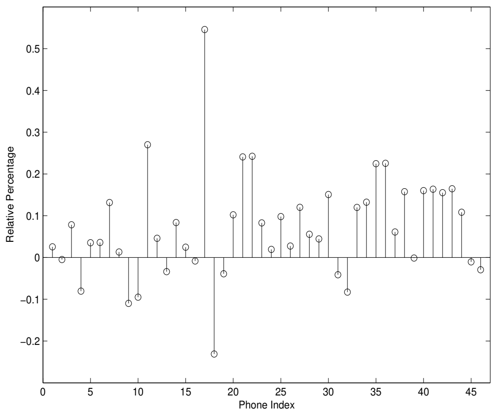

Figure 1: Difference in MRASTA and TANDEM phoneme confusion matrix.

3 The Dempster-Shafer Theory of Evidence

The Dempster-Shafer (DS) Theory of Evidence (see [6]) 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 $\sum_{A \cap B = \oslash} m_1(A) m_2 B < 1$. 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 beliefs.

In next section, we show how to transform NN outputs into belief functions and combine them according to DS rule.

4 Combination of NN outputs

We consider the output of a Neural Network trained to estimate posterior probability of a target class (i.e. a phoneme posterior) [15]. 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 $H = \sum_{i=1}^{k} p_i log(p_i)$. Ignorance is supposed to be total (i.e. $m(\Theta) = 1$) when entropy of the output achieves its maximum value $H_{max} = \sum_{i=1}^{k} \frac{1}{k} log(\frac{1}{k})$. Under those considerations a possible choice for a BPA is represented by:

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

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

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.

In [5], we investigated several BPAs; the most effective combines information about $m_i(\theta_i)$, $m_i(\neg\theta_i)$ and $m_i(\Theta)$.

We could define such a 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.

Let us consider now the case in which we have two different Neural Networks and their corresponding BPA obtained as described in previous section. Those BPA can now be combined applying orthogonal sum (4).

$$\begin{aligned} m(\theta_i) &= \{m_a(\theta_i)m_b(\theta_i) + m_a(\theta_i)m_a(\Theta) + \\ &+ m_b(\theta_i)m_a(\Theta)\}/Z \qquad (10) \\ m(\neg\theta_i) &= \{m_a(\Theta)m_b(\neg\theta_i) + m_b(\Theta)m_a(\neg\theta_i)\}/Z \qquad (11) \\ m(\Theta) &= \{m_a(\Theta)m_b(\Theta)\}/Z \qquad (12) \\ Z &= 1 - m_a(\neg\theta_i)m_b(\theta_i) - m_b(\neg\theta_i)m_a(\theta_i) \qquad (13) \end{aligned}$$

Combination rules (10 - 13) 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.

5 LVCSR experiments

The use of NN features resulted in considerable improvements in LVCSR system when used in combination with classical spectral features ([16]). We investigate here performances of combination of 9-frames PLP and MRASTA Neural Networks in transcription of meetings data. For investigation purposes, we run experiments without concatenation with other features and compare with PLP augmented with dynamic features.

The training data for this system comprises of individual headset microphone (IHM) data of four meeting corpora; the NIST (13 hours), ISL (10 hours), ICSI (73 hours) and a preliminary part of the AMI corpus (16 hours). Acoustic models are phonetically state tied triphones models trained using standard HTK maximum likelihood training procedures. The recognition experiments were conducted on the NIST RT05s [9] evaluation data. We use the reference speech segments provided by NIST for decoding. The pronunciation dictionary is same as the one used in AMI NIST RTO5s system [14]. Juicer large vocabulary decoder [17] was used for recognition with a pruned trigram language model.

	FER	Cross entropy
Sum	31.9%	$40 \cdot$ 1E+6
Prod	31.7%	$50 \cdot$ 1E+6
Inv.entropy	31.9%	$40\cdot$ 1E+6
DS	31.7%	$53 \cdot$ 1E+6

Table 1: FER and Cross entropy for different combination rules

We compare several combination rules i.e. sum rule, product rule, inverse entropy [4] and the Dempster-Shafer (DS) combination. In DS combination the value of γ in equation 7 is set heuristically and is further investigated in section 5.1. The experimental framework is the following one: two different neural network (9 frames PLP and MRASTA) are trained on all available data. Posterior

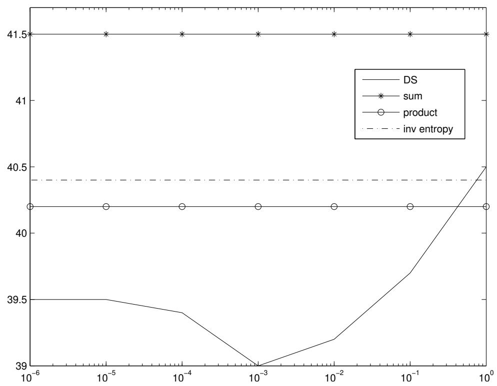

Figure 2: WER function of the heuristic factor γ for Dempster-Shafer combination. Constant lines are WER for Sum, Product and Inverse entropy combination rules.

Features	TOT	AMI	CMU	ICSI	NIST	VT
PLP+D+A	42.4	42.8	40.5	31.9	51.1	46.8
TANDEM	46.6	41.4	43.7	31.3	54.5	64.9
MRASTA	45.9	48.0	41.9	37.1	54.4	48.8
Sum	41.5	41.1	37.6	30.4	50.2	49.8
Prod	40.2	39.4	37.1	29.5	48.4	47.7
inventropy	40.4	39.8	37.0	29.6	48.3	48.7
DS combination	39.0	39.0	36.8	28.1	45.8	46.6

Table 2: WER for RT05 evaluation data.

distribution are combined according to previous rules and Log/KLT transform is estimated for each combination. Table 2 reports WER for different combination rules on RT05 evaluation data.

Out of the four rules, DS combination is giving the best result in all subset of evaluation data. The case of VT data is interesting case in which a feature stream (9 frames PLP) has very poor performances (because of noise) while MRASTA shows a performance comparable to PLP features. Ideally, results of combined feature streams should be better than result of the best individual stream. This is not verified for the sum and inverse entropy combination in which the use of two streams is not giving any further improvements. On the other hand, product rule gives 1% absolute improvement w.r.t MRASTA and Dempster-Shafer rule gives 2% absolute improvement.

5.1 Analysis of results

To better understand the effect of DS combination, it is useful to consider errors at the frame level for different combination rules. Table 1 provides Frame Error Rate (FER) for the combined features. Sum and Inverse entropy rules provides similar FER (31.9%) while Product and Dempster-Shafer both hold a FER of 31.7%. Differences between combination rules are not significant and cannot explain the difference in performances.

Anyway NN based features are not used for making classification at frame level but for generating an intermediate probabilistic representation of the data. For this reason, it is rather useful to consider a metric that takes into account the probabilistic nature of the output. Let us consider the cross entropy defined as $\sum_i t_i log\, p_i$ where p_i is the ith output of the NN and t_i is one for the reference phoneme and zero for all other phonemes. Table 1 show as well values of Cross Entropy (CE) for different combination rules : DS rule achieves the highest CE followed by Product rule and Sum and Inverse entropy rules. This ordering is consistent with what observed in terms of WER. Thus we can conclude that compared to other rules, DS achieves a higher cross entropy in between phoneme posterior and reference target that translates in actual gain in Word Error Rate.

In equation 7, we defined in a heuristic way the transformation of a probabilistic output into a BPA. The main idea is to increase BPA of ignorance $m_i(\Theta)$ proportionally to entropy. Exponent factor γ is supposed to better fit this match in case of non-linear dependence. When $\gamma \to 0$, $\alpha \to 1$ i.e. the BPA for ignorance is zero. Thus it is important to investigate the impact of γ on final feature combination. Figure 2 plots WER with respect to the value of γ. When γ is in the range of zero and one DS combination is consistently outperforming all other proposed rules. On the other hand when γ is equal to one, DS performs worst than product and inverse entropy rules. This suggest that the relationship in between BPA and entropy is not linear.

Furthermore, analysis of phonetic confusion matrix for combined feature streams shows that *for each* of the phonetic targets, the combined features achieves higher classification accuracy than the single feature streams. Considering the motivation for combining classifiers described in section 2, we can conclude that those techniques are effective in reducing total Frame Error Rate, Cross Entropy and recognition accuracy for each target.

Those findings on LVCSR system are consistent with what previously observed on smaller databases [5].

6 Conclusion

In this paper, we investigate the use of Dempster-Shafer (DS) combination rule for multi-stream LVCSR system. DS rule is based on theory of evidence which generalizes classical probability theory. Main advantage of this framework is the explicit representation of ignorance which is estimated using entropy from a given feature stream. In our previous work, we showed that it outperforms classical combination rules like sum, product or inverse entropy on small vocabulary task. We consider here a LVSCR task for transcription of meetings data. On RT05 evaluation data DS combination outperforms by 1.4% absolute inverse entropy and by 1.2% absolute product rule. Dependency on heuristic function that transforms probabilities into beliefs is investigated as well.

7 Acknowledgments

This material is based upon work supported by the EU under the grant DIRAC IST 027787 and by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. 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). Authors thank Thomas Hain and AMI ASR team for their help with the meeting system.

References

[1] Bourlard H. and Dupont S., "A new asr approach based on independent processing and recombination 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] Misra H., Bourlard H., and Tyagi V., "Entropy-based multi-stream combination," in *Proceedings of ICASSP*, 2003.

[5] Valente F. and Hermansky H., "Combination of acoustic classifiers based on dempster-shafer theory of evidence," *Proc. ICASSP 2007*.

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

[7] 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.

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

[9] "http://www.nist.gov/speech/tests/rt/rt2005/spring/," .

[10] H. Bourlard and C.J. Wellekens, *Speech Pattern Discrimination and Multilayer Perceptrons*, Academic Press, 1989.

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

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

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

[14] Hain T. et al, "The 2005 AMI system for the transcription of speech in meetings," *NIST RT05 Workshop Edinburgh, UK.*, 2005.

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

[16] Zhu Q., Chen B., Morgan N., and Stolcke A., "On using mlp features in lvcsr," *Proceedings of ICSLP 2004*.

[17] Moore D et al., "Juicer: A weighted finite state transducer speech coder," *Proc. MLMI 2006 Washington DC*, 2006.