# A DEMPSTER-SHAFER BASED FUSION APPROACH FOR AUDIO-VISUAL SPEECH RECOGNITION WITH APPLICATION TO LARGE VOCABULARY FRENCH SPEECH

*S. Foucher, F. Laliberté, G. Boulianne, L. Gagnon*

R&D Department, CRIM, 550 Sherbrooke Street West, Suite 100, Montreal (QC), CANADA, H3A 1B9
email: {samuel.foucher, france.laliberte, gilles.boulianne, langis.gagnon}@crim.ca

## ABSTRACT

This work explores a new way of fusing audio and visual information for audio-visual automatic speech recognition in the context of a large vocabulary application. Mouth shape information is extracted off-line and integrated into a speech recognition system using a phoneme-based Dempster-Shafer fusion approach. The fusion methodology assumes that the audio information about the phonemes is a precise Bayesian source while the visual information is an imprecise evidential source. This ensures the visual information does not degrade significantly the audio information in situation where the audio performs well in controlled noiseless environment. Bayesian and simple consonance belief structures are explored and compared, along with standard stack-based fusion.

## 1. INTRODUCTION

This paper deals with the problem of improving audio-only speech recognition performances using low level visual clues. The main objective is to decrease the Word Error Rate (WER) of a speech recognition system when audio acquisition conditions are poor and the resulting Signal to Noise Ratio (SNR) is low. Most of the recent Audio-Visual Automatic Speech Recognition systems (AV-ASR) can be split into two main fusion categories: (1) feature-based or stack vector-based fusion, where visual features are simply concatenated with the audio features and (2) decision based fusion, where fusion is performed at higher level [6][7]. Most works use linear representations and a strict Bayesian framework. Unlike those works, our approach makes use of a decision fusion process based on the Dempster-Shafer (DS) theory [8].

Visual evidence is highly imprecise compared to audio mainly because most of the articulators involved in speech production are not visible (tongue body, velum, glottis). Visual evidence is also impaired by coarser sampling rate (usually three audio frames for one visual), variations of the head pose and the mouth appearance. Visual and audio information are thus strongly heterogeneous data.

The DS theory [8] offers a very powerful framework for fusion of heterogeneous data. It is being applied to various data fusion problems [1][3]. In particular, an extension of this theory, called Transferable Belief Model (TBM), offers an even more flexible framework [3]. Recently, it has been proposed to use the DS theory to fuse decisions from an imprecise information source (modeled by an evidential mass function) with a precise source (modeled by a Bayesian probability) [1]. This approach has the advantage of producing a Bayesian mass function which can be further handled by standard Bayesian algorithms. Following this idea, we model the visual information by an evidential mass function while the audio information remains a Bayesian mass function. Some concepts of the TBM framework are exploited such as the ballooning extension and the conjunctive rule of combination [3].

The paper is organized as follows. Section 2 briefly exposes the visual learning approach which is based on the training of a set of binary classifiers using a Kernel Linear Discriminant Analysis (KLDA) [9]. Section 3 describes our DS-based fusion approach along with the visual information mass functions assignment we have tested; Bayesian and simple consonance. Section 3 also describes how the Bayesian audio phoneme likelihoods are merged with the visual phoneme evidence using the conjunctive TBM rule. Section 4 presents results on a large dataset of French Canadian broadcast news readings. Finally, we conclude and identify possible extension to our work.

## 2. VISUAL LEARNING

The visual learning aims at classifying mouth shapes related to particular phonemes or set of phonemes. Visual learning is difficult because of the high variability in the visual content. Mouth shape variations due to different phonemes are rather small compared to other sources of intra-subject variability like head pose variations, tracking errors and pronunciation variations. In the French speaking language, 36 phonemes are usually used. Clearly, distinct phonemes do not necessarily have corresponding distinct visual classes. Based on the mouth appearance only, phonemes are often grouped into few classes of visual phonemes called visemes. Single phonemes are not atomic representations and are heavily dependent on the context. Thus, groups of three phonemes (triphones) are usually considered in most speech

recognition algorithms [6]. However, due to practical limitations (e.g. not enough triphones in our database to train a triphone-based classifier; which are about tens of thousands for the French language), we retained the phoneme-based visual learning approach as a trade-off between the viseme and the triphone approaches, with the hope that our DS decision-based fusion algorithm can compensate by adapting the confidence level according to the context.

## 2.1. Visual Feature Formation

Visual feature extraction is based on an open source program developed by Intel and which is part of the library OpenCV [5]. The face and mouth detection are first done by a boosted cascade of classifiers using Haar-like features. Then, a Kalman filter tracks the detected mouth window from which we derive a greyscale vector of dimension 160 (10x16 pixels). In order to reduce the impact of lighting conditions, each feature was centered on the mean feature vector of each speaker. A projection was finally done onto a PCA subspace containing 95% of the total feature variance as a way of keeping a good learning robustness. The resulting low-level visual feature vector is of dimension 54.

## 2.2. Kernel Based Learning

Kernel-based learning techniques are efficient in describing high-dimensional non-linear manifold using Kernel PCA (KPCA) [9]. The kernel functions in KPCA allow for non-linear extensions of the linear feature extraction methods. The input vectors $\mathbf{x} \in \mathbb{R}^n$ are mapped into a new higher-dimensional feature space $\mathbf{F}$, using a mapping function $\phi : \mathbb{R}^n \rightarrow \mathbf{F}$, in which the linear methods are applied. One major drawback of Kernel-based learning is the size of the Gram matrix which increases in function of the number of training samples. Therefore, training on large datasets is prohibitive. In particular, multi-class KLDA with a large number of classes requires a large number of samples per class. Consequently, solving the generalized eigenvalue problem required by LDA is usually beyond actual computational power. We rather trained a set of pairwise binary classifiers instead of dealing directly with the multi-class problem. Once the pairwise classification is done we only need to combine $N = |\Omega^v| \times (|\Omega^v| - 1) / 2$ binary decisions to form a likelihood vector on the original set of visual classes $\Omega^v$. A simple voting algorithm is then used to build a histogram for each decision [4].

## 3. FUSION METHODOLOGY

### 3.1. Transferable Belief Model

The core element of the TBM is the basic belief assignment (or mass) function $m()$. The mass function assigns a belief on the subsets of the set $\Omega = \{\omega_i\}_{i=1}^M$ constituted by $M$ mutually exclusive hypotheses $\omega_i$. Based on the available evidence (the facts) $E$, the mass function is defined by

$$m^\Omega[E]() : 2^\Omega \rightarrow [0,1] \quad \text{with} \quad \sum_{B \subseteq \Omega} m^\Omega[E](B) = 1 \quad (1)$$

where $2^\Omega$ denotes the set of all subsets of $\Omega$ (the power set). In the DS theory, an additional normalization is imposed to ensure that the null hypothesis has a null belief ($m(\varnothing) = 0$). The TBM does not require this. We call focal set (noted $F$) the set of subsets of $2^\Omega$ having a non-null mass, i.e. $F = \{A \subseteq \Omega \mid m^\Omega[E](A) > 0\}$. The *belief* function *bel()* is defined as

$$bel : 2^\Omega \rightarrow [0,1] \text{ with } bel(A) = \sum_{\varnothing \neq B \subseteq A} m^\Omega[E](B), \forall A \subseteq \Omega \,(2)$$

The degree of belief *bel(A)* represents the amount of justified and specific degree of support to the hypothesis *A*. The *plausibility* function *pl(A)* is the degree of support that could be attributed to *A* but can also support another subset:

$$pl : 2^\Omega \rightarrow [0,1] \text{ with } pl(A) = \sum_{B \cap A \neq \varnothing} m^\Omega[E](B), \forall A \subseteq \Omega \,(3)$$

The two quantities *bel(A)* and *pl(A)* are often interpreted as a lower and upper bound of an unknown probability measure *P* on *A*. In addition, the difference *pl(A) - bel(A)* is an indicator of the degree of knowledge imprecision on *P(A)*.

### 3.1.1. The Ballooning Extension

The ballooning extension is a useful concept when belief is available on a subset $\Omega'$ of the full set of hypothesis $\Omega$. This happens, for instance, when beliefs are built on a limited set and one discovers afterwards that alternatives had not been considered. In particular, this extended mass function produces a new plausibility function:

$$pl^{\Omega' \Uparrow \Omega}(A) = \begin{cases} pl^{\Omega'}(A), & \text{if } A \subseteq \Omega \\ 1, & \text{if } A \subseteq \overline{\Omega'} \end{cases} \quad (4)$$

We use this principle to extend the visual beliefs on the phoneme subset $\Omega^v = \Omega \setminus \{G, N \sim\}$ to the full set of phonemes $\Omega$ ($\Omega^v \subseteq \Omega$). This is useful because the *G* and *N~* phonemes are rare in the French language and we do not have enough samples to properly learn them.

### 3.1.2. Mass Function Construction

As proposed in the statistical evidential theory [8], we simply construct mass functions from observed likelihoods

$\{p(x \mid \omega_i)\}_{i=1,\dots,M}$ . Different mass functions can be constructed depending on the type of partition on $\Omega$ [8][10]. The two limiting cases are:
- the simple consonant belief function [8]:

$$bel^{\Omega}[E](\omega_j) = 1 - \frac{\max\limits_{\omega_i \in \overline{\omega_j}} \{p(x \mid \omega_i)\}}{\max\limits_{\omega_i \in \Omega} \{p(x \mid \omega_i)\}}, \forall \omega_j \in \Omega \quad (5)$$

- the Bayesian belief function:

$$bel^{\Omega}[E](\omega_j) = \frac{p(x \mid \omega_j)}{\sum\limits_{k=1}^{M} p(x \mid \omega_k)}, \forall \omega_j \in \Omega \quad (6)$$

### 3.2. Application to AV-ASR

In the literature [6][7], the HMM state-dependent emission of an audio-visual observation vector is represented by a direct product of the probability for each audio frame *t* and the HMM context dependent state *c*:

$$P(\mathbf{o}_{av,t} \mid c) = P(\mathbf{o}_{a,t} \mid c)^{\lambda_{a,c,t}} P(\mathbf{o}_{v,t} \mid c)^{\lambda_{v,c,t}}, \forall c \in C \quad (7)$$

where $\lambda_{a,c,t}$ ($\lambda_{v,c,t}$) is the reliability factor and $\mathbf{o}_{a,t}$ ($\mathbf{o}_{v,t}$) are the observed low-level feature vectors for the audio (resp. visual) source. The non-negative reliability factors control each modality contribution. Usually, *c* is a context-dependent phoneme ($c = \{\omega_i, \omega_j, \omega_k\}$) and is modeled using a Markovian method. For the visual information, the likelihood is not contextual ($P(\mathbf{o}_{v,t} \mid c) = P(\mathbf{o}_{v,t} \mid \omega_j)$).

Here, we propose to formulate (7) within the evidential framework, i.e. using Dempster's rule of combination:

$$m^{\Omega}[a,v](c) = \left(m^{\Omega}[a] \otimes m^{\Omega}[v]\right)(c) \quad (8)$$

where $m^{\Omega}[a]$ (resp. $m^{\Omega}[v]$) is the mass function associated to the audio (resp. visual) information and $m^{\Omega}[a,v]$ is the combined audio-visual mass. We assume that the audio modality is a precise Bayesian source ($m^{\Omega}[a](c) = P(\mathbf{o}_{a,t} \mid c)$) so that the audio mass function is directly the observed audio likelihood. The audio-visual mass function (8) with the reliability coefficients becomes

$$m^{\Omega}[a,v](c) = P(\mathbf{o}_{a,t} \mid c)^{\lambda_a} \left[pl^{\Omega}[v](c)\right]^{\lambda_v} \quad (9)$$

Equation (9) can be seen as a generalization of (7). We still need to express the visual plausibility function for the chosen belief structures. We give here the results for the Bayesian and simple consonance. We assume known a set of hypothesis $\Omega^v = \{\omega^{(1)}, \dots, \omega^{(M)}\}$ resulting from the

ordering of the hypothesis $\omega_i \in \Omega$ according to a set of observed likelihoods $p^v(\omega^{(1)}) > p^v(\omega^{(2)}) > \dots > p^v(\omega^{(M)})$.

#### 3.2.1. Bayesian Mass Function
In this case, the focal set is $F = \Omega$ and the mass function is given by (6). Applying the ballooning extension (4) we get, for all $A \in \Omega$,

$$pl^{\Omega}[v](A) = \begin{cases} 1, & \text{if } A \in \overline{\Omega^v} \\ m^{\Omega^v}[v](A), & \text{if } A \in \Omega^v \end{cases} \quad (10)$$

#### 3.2.2. Simple Consonant Mass Function
In this case, the focal set is $F = \{\omega^{(1)}, \Omega\}$ and the visual mass function is given by (5):

$$m^{\Omega^v}[v](A) = \begin{cases} 1 - \dfrac{p^v(\omega^{(2)})}{p^v(\omega^{(1)})}, & A = \omega^{(1)} \\ \dfrac{p^v(\omega^{(2)})}{p^v(\omega^{(1)})}, & A = \Omega^v \end{cases} \quad (11)$$

After applying the ballooning extension we obtain, for all $A \in \Omega$, the visual plausibility function:

$$pl^{\Omega}[v](A) = \begin{cases} 1, & \text{if } A \in \omega^{(1)} \cup \overline{\Omega^v} \\ m^{\Omega^v}[v](\Omega^v), & \text{if } A \notin \omega^{(1)} \cup \overline{\Omega^v} \end{cases} \quad (12)$$

### 4. RESULTS

Our speech corpus consists of 740 distinct television news utterances read by 26 native French Canadian speakers. The AV data were collected in-house and totalize 4.5 hours of upper body frontal color video of 380x540 pixels at 30 fps. The audio data is sampled at 16 kHz and linearly quantized to 16 bits with a SNR=26dB. The first 10 MFCC are extracted every 100 ms and combined with their first- and second- time-derivative to form an audio feature vector of dimension 30. For stack-based fusion, second-time derivatives are replaced by 10 visual features. For each subject, 20% of the data was kept as test and 80% as training. A subset of the test set, but limited to 6 subjects, was also used as a development set to tune the training parameters. Training was done using clean audio data within a multi-speakers and speaker-independent framework, i.e. the same subjects were used in training and testing but training and testing utterances were distinct. Noisy versions of the test data were generated by adding speech babble. The fusion of probabilities along the lines described in Section 3.2 was done within the CRIM's speech recognition system which is based on Hidden Markov Models (HMM), Gaussian mixtures, and N-gram language model [2][11].

Note that results are reported for recognition in *unmatched* conditions, since models were trained on clean data only. WER performance on the test set for the clean (26 dB) and noisy data (18.9 dB, 13.8 dB, 9.55 dB) are given on Figure 1. Four types of fusion approaches have been tested: (1) decision-based simple consonance, (2) decision-based Bayesian, (3) stack-based (with KLDA components), and (4) combination of stack- and decision-based simple consonance. The best results have been obtained with the last approach giving a WER reduction from 82% to 67% at 9.55 dB noise, or an equivalent gain of 4 dB for the SNR.
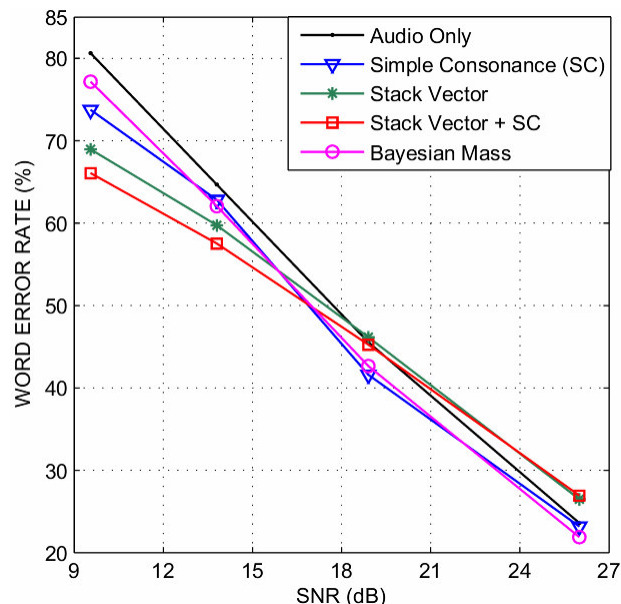


**Figure 1:** Average WER on test data for babble noise

It is difficult to compare our results to other published works as the application area (large French vocabulary of high perplexity) and/or the dataset are different. In [6][7], the authors report WER reduction on a large vocabulary English recognition application from 48% (audio only) to 35% (with AV fusion) at 10 dB noise. We get a similar relative reduction in our case (81% to 65%). This is rather encouraging taking into account that (1) we have used simple, and maybe not optimal mass functions and (2) did not take into account the contextual visual information. Regarding the higher WER we have for the clean data, one can say that this value is very dependent on the dataset and the acoustic and language models. Furthermore, we did not use any complex data pre-processing like MLT, KLDA or clean/noisy data mismatch correction technique in our speech recognition system.

## 5. CONCLUSION

We have reported about the use of DS and TBM probability theories as fusion approach for AV-ASR of large vocabularies, with application to French Canadian speech.

The evidential framework, and in particular the TBM theory, offers many advantages: (1) possibility to extend mass functions to a larger set of hypothesis (ballooning extension), (2) manipulation of non-singleton hypothesis, and (3) modeling of the global imprecision. We have chosen a statistical evidence framework mainly because it produces simple and efficient mass functions that are easily combinable. This leads to performance results on WER that are encouraging and comparable to other published results.

To our knowledge, our work is the first to address AV-ASR of French Canadian speech. It also deals with a large vocabulary application which is still an open research issue for any language. As a contribution to this field, we will soon release on the Web our audio-visual database for research activities.

## REFERENCES

[1] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, W. Pieczynski, "Dempster-Shafer Fusion in Markov Fields Context", *IEEE Trans. on GRS*, Vol. 39, No 8, pp. 1789-1798, 2001.

[2] G. Boulianne, J. Brousseau, P. Ouellet, P. Dumouchel, "French Large Vocabulary Recognition with Cross-Word Phonology Transducers", *Proc. ICASSP-2000*.

[3] F. Delmotte, P. Smets, "Target Identification Based on the Transferable Belief Model Interpretation of Dempster-Shafer Model". *IEEE Trans. on SMC*. A:34, pp. 457-471, 2004

[4] J. Fürnkranz, "Round Robin Classification", *Journal of Machine Learning Research 2*, pp. 721-747, 2002.

[5] A.V. Nefian, L. Hong Liang, Xiao X. Liu, X. Pi, C. Mao, K. Murphy, "A Coupled HMM for Audio-Visual Speech Recognition", *Proc. ICASSP-2002*.

[6] G. Potamianos, C. Neti, J. Luettin, I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview", *In: Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, P. Perrier (Eds.), MIT Press (In Press), 2004.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, "Automatic Recognition of Audio-Visual Speech: Recent Progress and Challenges", *Proceedings of the IEEE*, vol. 91, no. 9, Sep. 2003.

[8] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, Princeton NJ, 1976.

[9] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[10] P. Walley, "Belief Function Representation of Statistical Evidence". *The Annals of Statistics*, 15(4):1439-1465, 1987.

[11] L. Gagnon, S. Foucher, F. Laliberté, G. Boulianne, C. Chapdelaine, "Exploration de la reconnaissance audio-visuelle du français québécois pour une application à large vocabulaire", Technical report, CRIM-05/06-08, June 2005 (in French)