FEATURE SPACE MUTUAL INFORMATION IN SPEECH-VIDEO SEQUENCES

Torsten Butz and Jean-Philippe Thiran

Swiss Federal Institute of Technology (EPFL)
Signal Processing Institute (ITS)
CH-1015 Lausanne, Switzerland
WWW: http://ltswww.epfl.ch, FAX: +41-21-693-7600
{torsten.butz,jp.thiran}@epfl.ch

ABSTRACT

We present an approach to directly study mutual relationships between audio and video signals for multimedia applications. The presented approach is mathematically based on information theory and is closely related to information theoretic classification. We show that very simple features of the audio- resp. video-channel can already contain lots of mutual information between both modalities. The mathematical approach is very general though and not restricted to the presented multimedia application.

1. INTRODUCTION

Classically digital signal processing has treated different signals as quasi-independent entities. The natural connection of *multi*-modal signals coming from the same physical scene has mostly been neglected. Lately though, several attempts have been made to explore their relationship for signal processing algorithms. Examples for multimedia applications are shown in [1] and [2]. In [3], we have presented a very general information theoretic approach to handle multi-modal signals. There is a vast number of at first sight completely unrelated applications that could benefit from our approach.

In this paper we want to apply this theory specifically to multimedia signals. A general and widely applicable multimodal approach can have an impact on a broad range of applications in this field. For example human-computer interfaces [4], speaker recognition [5] or media conversion could profit from exploring simultaneously the audio and video signal. In particular, we show that our theory [3], which also generalizes multi-modal medical image registration with mutual information [6], [7], is general enough to enhance speaker detection in audio-video sequences.

First we shortly summarize the information theoretic approach of feature space mutual information for multi-modal signal processing. Afterwards we show how it can be directly applied to speaker detection.

2. FEATURE SPACE MUTUAL INFORMATION

Speaker detection attempts to detect a speaker in a video sequence. The physical relation between the speaker's mouth motion and the resulting speech lets us expect to find a relationship between the corresponding digital signals. Therefore we wanted to detect the pixels in the video that carry most information about the audio signal. Intuitively one might calculate directly the mutual information over time between each video pixel and the audio signal. The maximum should lie at the speaker's mouth and therefore detect the region that is physically responsible for the formation of the audio. Unfortunately it's not obvious that simply the image intensities of the mouth would carry this information. In fact we will show that they do not. But we will determine an other video feature that turned out to be very appropriate for our task.

But first of all, we want to give a mathematical explanation to use mutual information for speaker detection. Furthermore we show that maximum mutual information naturally incorporates optimal feature selection/extraction. We called the resulting measure "Feature Space Mutual Information", reflecting the simple fact that we calculate this statistical pseudo-distance measure on features extracted from the initial signals and not on the signals themselves.

2.1. Fano's Inequality for Multi-modal Signals

Fig. 1 shows how it is possible to formally connect multimodal signals through a joint probability distribution of their extracted features. We used joint histogramming for density estimation [8]. This lets us easily construct two related Markov chains from the multi-modal signals:

$$S \rightarrow V \rightarrow F_V \rightarrow F_A^{est} \rightarrow A^{est} \rightarrow S^{est} \qquad (1)$$

and

$$S \rightarrow A \rightarrow F_A \rightarrow F_V^{est} \rightarrow V^{est} \rightarrow S^{est}, \qquad (2)$$

0-7803-7304-9/02/$17.00 C2002 IEEE

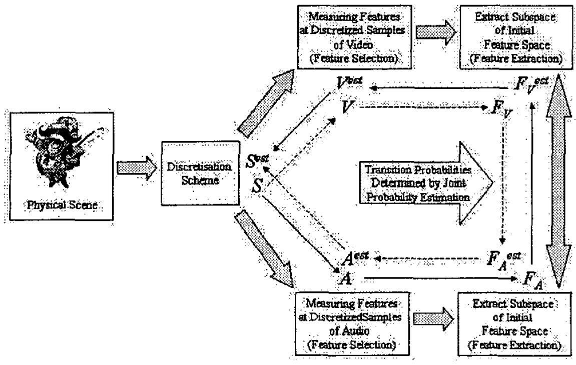

Measuring Features at Discretized Samples of Video (Feature Selection)
Extract Subspace of Initial Feature Space (Feature Extraction)
Physical Scene
Discretisation Scheme
S^{est}
S
V^{est}
V
F_V^{est}
F_V
Transition Probabilities Determined by Joint Probability Estimation
A^{est}
A
F_A^{est}
F_A
Measuring Features at DiscretizedSamples of Audio (Feature Selection)
Extract Subspace of Initial Feature Space (Feature Extraction)

Figure 1: It's possible to construct Markov chains from multi-modal signals, where the joint histogram between the final features (F_V and F_A) is the connecting block between the audio and video signal.

where S is a uniform random variable (RV) that generates the sampling positions in the initially continuous signals. V and A are the RVs modeling the specific measurements at the positions generated from S. What exactly is measured is a feature selection step. The measured features V and A can be multi-dimensional which might require next a feature extraction step. The output after feature extraction is modeled by the RVs F_V and F_A respectively. Furthermore we need a probability estimation to estimate the transition probabilities from F_V to F_A^{est} and from F_A to F_V^{est}. Both probabilities can be estimated from the joint probability distribution of F_V and F_A. We used joint histogramming to approximate this probability density [8].

For both Markov chains, it's possible to estimate a lower bound of the error probability $P_{e1} = Pr(S \neq S^{est})$ resp. $P_{e2} = Pr(S \neq S^{est})$, that S^{est} does not equal the initial input S to the chains. These bounds can be derived using Fano's inequality [9] and several times the data-processing inequality [10]. As shown in [3], the lower bound of the error probability P_{e1} of the Markov chain of eq. 1 can be derived as follows:

$$\begin{aligned} P_{e1} &= Pr(S \neq S^{est}) && (3)\\ &\geq \frac{H(S|A^{est})-1}{\log|\Psi|} && (4)\\ &= \frac{H(S)-I(S,A^{est})-1}{\log|\Psi|} && (5)\\ &= 1-\frac{I(S,A^{est})+1}{\log|\Psi|} && (6)\\ &\geq 1-\frac{I(F_V,F_A)+1}{\log|\Psi|}. && (7) \end{aligned}$$

Analogously, we have for the Markov chain of eq. 2:

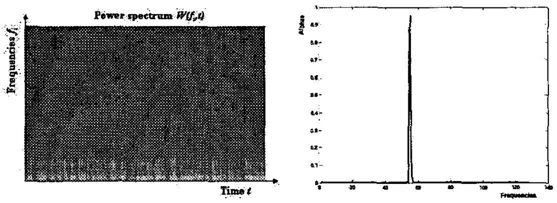
$$\begin{aligned} P_{e2} &= Pr(S \neq S^{est}) && (8)\\ &\geq \frac{H(S|V^{est})-1}{\log|\Psi|} && (9)\\ &= \frac{H(S)-I(S,V^{est})-1}{\log|\Psi|} && (10)\\ &= 1-\frac{I(S,V^{est})+1}{\log|\Psi|} && (11)\\ &\geq 1-\frac{I(F_A,F_V)+1}{\log|\Psi|}. && (12) \end{aligned}$$

$I(.,.)$ stands for the Shannon mutual information and $|\Psi|$ is the number of possible measurement positions that can be generated from the RV S. Eq. 5 and 10 follow directly from Fano's inequality and the definition of Shannon's mutual information. For eq. 6 and 11, we used the fact that S is a uniform RV and therefore has entropy $\log|\Psi|$. Using the data-processing inequality, we find directly the weakened bounds of eq. 7 and 12. By symmetry of mutual information, the final lower bounds of P_{e1} and P_{e2} are equal.

2.2. Audio-video Signals

For audio-video signals we want to find the features in the audio and video signal that minimize the presented lower bounds on the error probabilities of eq. 7 and 12 in the region of the speaker's mouth. From eq. 7 and 12 it follows that this is equivalent to having a large feature space mutual information $I(F_V, F_A)$ in this area. On the other hand a large bound should result in the regions where the movements are not caused by the speaker's lips and are therefore unrelated to the speech signal. So that's where $I(F_V, F_A)$ should be small.

To represent the information of the audio signal, we first converted it into a power-spectrum (fig. 2a).



(a) (b)

Figure 2: a) The power spectrum of the video sequence. At each time point we have the power coefficients of several frequencies. b) The alphas for which the weighted sum of eq. 15 has maximum entropy.

In order to deal with this multi-dimensional audio signal, we included a linear feature extraction step in the algorithm. As for any couple of RVs X and Y, we have $H(X) \geq I(X,Y)$ and from eq. 7 and 12 we get a weakened lower bound for the error probabilities $P_{\{e_1,e_2\}}$:

$$\begin{aligned} P_{\{e1,e2\}} &\geq 1 - \frac{I(F_V,F_A)+1}{\log|\Psi|} \\ &\geq 1 - \frac{H(F_A)+1}{\log|\Psi|}. \end{aligned} \tag{13}$$

Therefore we looked for the linear combination of the power spectrum coefficients $W(f_i,t)$ (fig. 2a) that carries most entropy:

$$\vec{\alpha}^{opt} = \arg \max_{\vec{\alpha}:|\vec{\alpha}|=1,\alpha_i \geq 0} H(\sum_i \alpha_i \cdot W(f_i,t)). \tag{14}$$

Detailed information about the maximum entropy principle can be found in [11], [12]. The finally obtained audio-feature is therefore defined by

$$F_A(t) = \sum_i \alpha_i^{opt} \cdot W(f_i,t). \tag{15}$$

In fig. 2b, we show for one sequence the weights α_i^{opt} that maximize the entropy of eq. 15 and therefore define the audio-features F_A of the audio signal.

Several other audio-features could be imagined. In particular we could learn the features that have effectively the highest mutual information with the video signal in the region of the speaker's mouth. In that case we would directly take profit from eq. 7 and 12 for the feature extraction step.

3. RESULTS

We want to show two important points about the presented theory. First of all that there exist features that relate the mouth movements of a speaker directly to the corresponding speech signal. On the other hand we want to show that the choice of a particular feature representation is very crucial for the performance of the algorithm. There are features that contain lots of information (have lots of entropy), but are unrelated to the other signal. Other features represent this dependency much better and result in very good results.

The straight forward approach to quantify the dependency (in the sense of eq. 7 and 12) between an audio and video signal of a speaker would be to calculate for each pixel the mutual information between its intensities and the audio-feature of eq. 15. In fig. 3 we show the corresponding results.

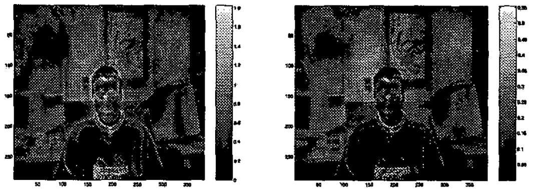

(a) (b)

Figure 3: a) We show the intensity entropies for each pixel of the sequence. It shows that our sequence contained lots of motion in the background of the scene (people passing, waving arms, etc.). b) The mutual information between the pixel intensities and the audio-feature has been calculated. We see that there is not a particularly high mutual information in the region of the speaker's mouth.

We can see that this straight forward approach doesn't lead to the result we could have expected. It seems that the pixel intensities of the speaker's mouth don't carry much information about the audio signal. Instead we propose a local feature that is more related to intensity changes than to the intensities themselves:

$$F_V(i,j,t) = \sum_{l,m=-1}^{1} g_{t+1}(i+l,j+m) - g_{t-1}(i+l,j+m), \tag{16}$$

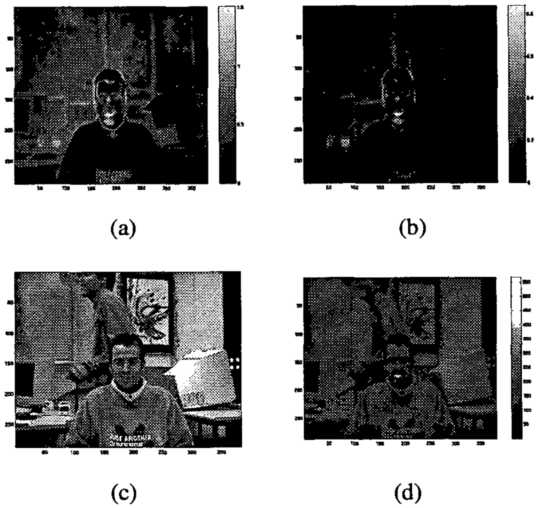
where $g_t(i,j)$ stands for the intensity of a pixel at coordinates (i,j) in the frame at time t.

Thereafter we calculated for each pixel in the scene the mutual information between the resulting audio- and video-feature $I(F_V,F_A)$. As shown in fig. 4 a clear relationship between the speech and the speaker's mouth is obtained.

4. DISCUSSION

In fig. 4 we showed that the proposed approach, which estimates for each pixel in the sequence the feature space mutual information between the video-feature F_V (eq. 16) and the audio-feature F_A (eq. 15), nicely detects the speaker's mouth in the video scene.

The feature selection/extraction itself teaches us several interesting facts about the presented audio-video signal. From fig. 2b, we can see that just a very small band of the power spectrum carries most of the information about the speech signal. Furthermore the video features tell us that image intensities are too sensitive to small illumination changes in the sequence. If we use a more local feature such as the presented intensity differences between consecutive frames (eq. 16), we significantly improve the result. Furthermore small camera movements (jitters) can be compensated by averaging over a small region around the pixels.



(a) (b)

(c) (d)

Figure 4: a) shows the entropy of the video-features F_V for each pixel in the video scene. b) relates this video information to the extracted audio-features F_A of eq. 15 by calculating the feature space mutual information $I(F_V, F_A)$ for each pixel. c) shows a typical frame of the sequence. d) is simply the thresholded image of b) super-posed on the frame of c). It shows that the mutual information maxima lie clearly at the speaker's mouth.

5. CONCLUSION

We applied the general framework of feature space mutual information to speech-based speaker detection. The choice of the right features has shown to be very crucial for the quality of the results. The theory is general enough to incorporate feature learning algorithms to optimize the performance of information theoretic multimedia signal processing. In particular feature selection and extraction from the initial audio and video signal can be included easily in the algorithms.

The resulting features can reveal important information about how to handle particular multi-modal signals. For example we have shown that local intensity changes in the video sequence are more related to the resulting speech-signal than the intensities themselves. This also re-confirms the fact that lip motion carries very significant information about the resulting speech signal.

6. REFERENCES

[1] J.W. Fisher III, T. Darrell, W.T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems, Denver, USA*, November 2000.

[2] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo New York*, July 2000.

[3] T. Butz and J.-Ph. Thiran, "Multi-modal signal processing: An information theoretical framework," Tech. Rep. 02.01, Signal Processing Institute (ITS), Swiss Federal Institute of Technology (EPFL), 2002, http://ltswww.epfl.ch/~brain/.

[4] V. Pavlovic, A. Garg, J.M. Rehg, and T.S. Huang, "Multimodal speaker detection using error feedback dynamic bayesian networks," in *IEEE Conference on Computer Vision and Pattern Recognition, USA*, June 2000.

[5] S. Furui, "An overview of speaker recognition technology," in *In Proc. ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, 1994, pp. 1–9.

[6] W.M. Wells III, P. Viola, H. Atsumi S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, March 1996.

[7] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, April 1997.

[8] L. Devroye and L. Györfi, *Non-parametric Density Estimation*, John Wiley & Sons, 1985.

[9] R.M. Fano, *Transmission of Information: A Statistical Theory of Communication*, MIT Press and John Wiley & Sons, 1961.

[10] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[11] E.T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, 1957.

[12] J.C. Principe, D. Xu, and J.W. Fisher III, "Learning from examples with information theoretic criteria," in *Multimedia Signal Processing*, 2000.