Adaptive Multimodal Fusion by Uncertainty Compensation

Vassilis Pitsikalis, Athanassios Katsamanis, George Papandreou, and Petros Maragos

National Technical University of Athens, School of ECE, 15773, Athens, Greece
E-mail: {vpitsik, nkatsam, gpapan, maragos}@cs.ntua.gr

Abstract

While the accuracy of feature measurements heavily depends on changing environmental conditions, studying the consequences of this fact in pattern recognition tasks has received relatively little attention to date. In this work we explicitly take into account feature measurement uncertainty and we show how classification rules should be adjusted to compensate for its effects. Our approach is particularly fruitful in multimodal fusion scenarios, such as audio-visual speech recognition, where multiple streams of complementary time-evolving features are integrated. For such applications, provided that the measurement noise uncertainty for each feature stream can be estimated, the proposed framework leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We further show that previous multimodal fusion methods relying on stream weights fall under our scheme under certain assumptions; this provides novel insights into their applicability for various tasks and suggests new practical ways for estimating the stream weights adaptively. The potential of our approach is demonstrated in audio-visual speech recognition using either synchronous or asynchronous models.

Index Terms: multimodal fusion, audiovisual speech recognition, uncertainty compensation, Active Appearance Models, product HMMs, stream weights

1. Introduction

Motivated by the multimodal way humans perceive their environment, complementary information sources have been successfully utilized in many pattern recognition tasks. Such a case is audio-visual speech recognition (AV-ASR) [1], where fusing visual and audio cues can lead to improved performance relatively to audio-only recognition, especially in the presense of audio noise.

However, successfully integrating heterogeneous information streams is challenging. Different streams provide complementary information and multimodal schemes should properly elevate the discriminative abilities of each of the modalities. Such schemes should adapt to the effective environmental conditions, which can dissimilarly affect the reliability of the separate modalities by contaminating feature measurements with noise. For example, the visual stream in AV-ASR should be discounted when the visual front-end loses track of the speaker's face.

A common theme in many stream integration methods is the utilization of stream weights to equalize the different modalities. These operate as exponents to each stream's probability density and have been employed in fusion tasks of different audio streams [2] and audio-visual integration [3]. Although stream weighting has indisputable benefits as shown experimentally, it requires determining the weigths for the different streams; although various methods have been proposed for this purpose [4], a rigorous approach to dynamically adapt the stream weights is still missing.

We choose to explicitly take observation uncertainty of the different modalities into account. Modeling observation noise has been proven fruitful for single modality ASR [5, 6], and has been further pursued for applications such as speaker verification [7], multi-band ASR [2], and recently in enhancement techniques based on clean speech and noise modeling [8]. In our work, given an estimate of the feature measurement uncertainty, we show in a rigorous probabilistic framework how the models used for classification should be adjusted to compensate for this effect. The proposed scheme leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We demonstrate that previous stream weight-based multimodal fusion formulations are derived from our scheme under certain assumptions; this unveils their probabilistic underpinnings and provides novel insights into their applicability for various tasks. In this context, we further suggest new practical ways for estimating stream weights adaptively. Evaluating our method in AV-ASR experiments utilizing multistream HMMs improves their performance. We have also applied the proposed technique in conjuction with Product HMMs (P-HMM) [9, 10], which account for cross-modal asynchrony, obtaining promising results.

2. Feature Uncertainty, Adaptive Compensation, and Multimodal Fusion

Let us consider a pattern classification scenario. We measure a property (feature) of a pattern instance and try to decide to which of N classes $c_i, i = 1 \dots N$ it should be assigned. The measurement is a realization x of a random variable X, whose statistics differ for the N classes. Normally, for each class we have trained a model that captures these statistics and represents the class-conditional probability distributions $p_X(x|c_i)$. Our decision is then based on a proper rule, *e.g.* the Maximum A Posteriori (MAP) criterion: $\hat{c} = \operatorname{argmax} P(c_i|x) = \operatorname{argmax} p_X(x|c_i) \cdot P(c_i)$.

One may identify three major sources of uncertainty that could perplex classification. First, *inherent model ambiguity* due to improper modeling or limited discriminability of the feature set for the classification task. For instance, visual cues cannot discriminate between members of the same viseme class (e.g. /p/, /b/) [1]. Second, *parameter estimation uncertainty* that mainly originates from insufficient training. Use of the Bayesian Predictive Classification rule can possibly alleviate it [11]. Third, *observation uncertainty* due to errors in the measurement process or noise contamination. This is the type of uncertainty we mainly address in this paper.

Our work is supported by the European research program 'HIWIRE', by the European Network of Excellence 'MUSCLE', and in part by the European research program 'ASPI' and the NTUA research program 'Protagoras'.

2.1. Adaptive Compensation

We may represent observation uncertainty as a random variable E independent of any class c_i. For simplicity, it is regarded to be an additive Gaussian variable with probability distribution $p_E(e) = N(e; \mu_e, \Sigma_e)$. In this case, the measurement y, a realization of the random variable Y, is a noisy version of the clean feature X:

$$Y = X + E \tag{1}$$

So, it would be desirable to use the distributions $p_Y(\cdot|c_i)$ in order to account for observation uncertainty. However, we only have $p_X(\cdot|c_i)$ available.

To determine these distributions we assume that X and E are independent. Then the probability $p_Y(y|c_i)$ of the uncertain observation y given the class c_i may be expressed as convolution of $p_X(x|c_i)$ and $p_E(e)$: $p_Y(y|c_i) = \int_{-\infty}^{\infty} p_X(x|c_i)p_E(y-x)\,dx$. If $p_X(x|c_i) = N(x; \mu_i, \Sigma_i)$, then $p_Y(y|c_i)$ is also normal, namely

$$p_Y(y|c_i) = N(y; \mu_i + \mu_e, \Sigma_i + \Sigma_e), \tag{2}$$

indicating that it is simple to compensate for the observation uncertainty; the variances Σ_i of the trained models, namely the class-conditional probability distributions of the clean training data, should be adjusted by adding the variance Σ_e of the measurement noise. The means should be appropriately shifted as well. A similar approach has been previously followed in [6, 7, 8].

To further illustrate this point, we discuss how observation uncertainty influences decision in a simple 2-class classification task. The two classes are modeled by 2D spherical Gaussian distributions, $N(\mu_1, \sigma_1^2 I)$, $N(\mu_1, \sigma_2^2 I)$ and they have equal prior probability. If our observation y contains zero mean spherical Gaussian noise with covariance matrix $\sigma_e^2 I$ then the modified decision boundary is described by $\log N(y; \mu_1, \sigma_1^2 I + \sigma_e^2 I) - \log N(y; \mu_2, \sigma_2^2 I + \sigma_e^2 I) = 0$. In the case when σ_e^2 is zero, the decision should be made as in the clean case. If σ_e^2 is comparable to the variances of the models then the modified boundary significantly differs from the original one. So, neglecting uncertainty in the decision may easily lead to misclassifications. As uncertainty increases, decision becomes even more difficult since the observation is even less informative. For infinite uncertainty we have just to pick the class whose mean is closer to the observation, which is also intuitively expected, as demonstrated in Fig. 1.

2.2. Multimodal Fusion

For many applications one can get improved performance by exploiting complementary features, stemming from a single or multiple modalities. Let us assume that one wants to integrate S information streams which produce feature vectors $x_s, s = 1, \ldots, S$. If the features are statistically independent given the class label c, the conditional probability of the full observation vector $x_{1:S} \equiv (x_1; \ldots; x_S)$ is given by the product rule; application of Bayes' formula yields the class label probability given the features:

$$p(c|x_{1:S}) \propto p(c) \prod_{s=1}^{S} p(x_s|c). \tag{3}$$

This probability can then be used in classification, *e.g.* by the MAP rule $\hat{c} = \text{argmax}_{c \in \mathcal{C}}\, p(c|x_{1:S})$.

In an attempt to improve classification performance, several authors have introduced stream weights w_s as exponents in (3), resulting to the modified score

$$b(c|x_{1:S}) = p(c) \prod_{s=1}^{S} p(x_s|c)^{w_s}, \tag{4}$$

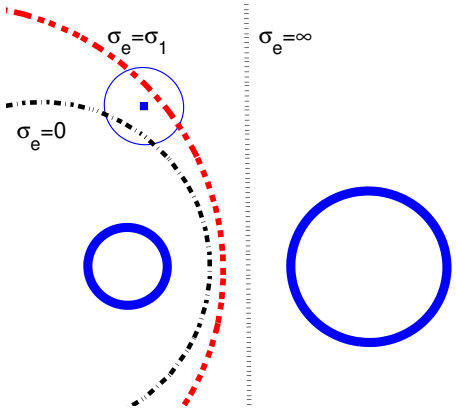

Figure 1: Decision boundaries for classification of a noisy observation (square marker) in two classes, shown as circles, for various observation noise variances. Classes are modeled by spherical Gaussians of means μ_1, μ_2 and variances $\sigma_1^2 I$, $\sigma_2^2 I$ respectively. The decision boundary is plotted for three values of noise variance (a) $\sigma_e = 0$, (b) $\sigma_e = \infty$, and (c) $\sigma_e = \sigma_1$.

which can also be seen in a logarithmic scale as a weighted average of individual stream log-probabilities. Such schemes have been motivated by potential differences in reliability among different information streams, and larger weights are assigned to information streams with better classification performance. Using such weighting mechanisms has experimentally been proven beneficial for feature integration in both intra-modal (*e.g.* multiband audio [2]) and inter-modal (*e.g.* audio-visual speech recognition [9, 4, 12]) scenarios.

However we find the stream weights formulation unsatisfactory in many respects. From a theoretical viewpoint, the weighted score b in (4) ceases having the probabilistic interpretation of (3) as class probability given the full observation vector $x_{1:S}$. Therefore it becomes unclear how to conceptually define, let alone implement, standard probabilistic operations, such as integrating-out a variable x_s (in the case of missing features), or conditioning the score on some other available information. From a more practical standpoint, it is not straightforward how to optimally select stream weights. Most authors set them discriminatively for a given set of environment conditions (*e.g.* audio noise level in the case of audio-visual speech recognition) by minimizing the classification error on a held-out set, and then keep them constant throughout the recognition phase. However, this is insufficient, since attaining optimal performance requires that we dynamically adjust the share of each stream in the decision process, *e.g.* to account for visual tracking failures in the AV-ASR case. Although there have been some efforts towards dynamically adjustable stream weights [4], they are not rigorously justified and are difficult to generalize.

We will now show that our approach for model adjustment in the presence of feature uncertainty naturally leads to a novel adaptive mechanism for fusion of different information sources. Since in our stochastic measurement framework we do not have direct access to the features x_s, our decision mechanism depends on the noisy version $y_s = x_s + e_s$ of the underlying quantity. The probability of interest is thus obtained simply by applying the convolution rule of Sec. 2.1 to each the independent stream separately. In the common case that the clean feature emission probability is modeled as a mixture of gaussians (MOG), *i.e.* $p(x_s|c) = \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N(x_s; \mu_{s,c,m}, \Sigma_{s,c,m})$, and the observa-

tion noise at each stream is considered gaussian, *i.e.* $p(y_s|x_s) = N(y_s; x_s + \mu_{e,s}, \Sigma_{e,s})$, it directly follows that

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \sum_{m=1}^{M_{s,c}} \rho_{s,c,m} N(y_s; \mu_{s,c,m} + \mu_{e,s}, \Sigma_{s,c,m} + \Sigma_{e,s}) \quad (5)$$

which simply means that we can proceed by considering our features y_s clean, provided that we shift the model means by $\mu_{e,s}$ and increase the model covariances $\Sigma_{s,c,m}$ by $\Sigma_{e,s}$. Note that, although the measurement noise covariance factor $\Sigma_{e,s}$ of each stream is the same for all classes c and all mixture components m, noise particularly affects the most peaked mixtures, for which the measurement noise uncertainty represented by $\Sigma_{e,s}$ is substantial relative to the modeling uncertainty due to $\Sigma_{s,c,m}$.

Although Eq. (5) is conceptually simple and easy to implement, given a good estimate of the measurement noise variance $\Sigma_{e,s}$ of each stream, it actually constitutes a highly adaptive rule for multisensor fusion. To appreciate this, and also to show how our scheme is related to the stream weights formulation of Eq. (4), we examine a particularly illuminating special case of our result. We make two simplifying assumptions:

1. The measurement noise covariance is a scaled version of the model covariance, *i.e.* $\Sigma_{e,s} = r_{s,c,m}\Sigma_{s,c,m}$ for some positive constant $r_{s,c,m}$ interpreted as the relative measurement error.
2. For every stream observation y_s the gaussian mixture response of that stream is dominated by a single component m_0 or, equivalently, there is little overlap among different gaussian mixtures.

Under these conditions the Gaussian densities in Eq. (5) can be approximated by $N(y_s; \mu_{s,c,m_0} + \mu_{e,s}, (1 + r_{s,c,m_0})\Sigma_{s,c,m_0})$; using the power-of-gaussian identity $N(x; \mu, \Sigma)^w \propto N(x; \mu, w^{-1}\Sigma)$ yields

$$p(c|y_{1:S}) \propto p(c) \prod_{s=1}^{S} \left[\tilde{\rho}_{s,c,m_0} N(y_s; \mu_{s,c,m_0} + \mu_{e,s}, \Sigma_{s,c,m_0}) \right]^{w_{s,c,m_0}} \quad (6)$$

where

$$w_{s,c,m_0} = 1/(1 + r_{s,c,m_0}) \quad (7)$$

is the *effective stream weight* and $\tilde{\rho}_{s,c,m_0}$ is a properly modified mixture weight which is independent of the observation y_s. Note that these effective stream weights are between 0 (for $r_{s,c,m_0} \gg 1$) and 1 (for $r_{s,c,m_0} \approx 0$) and discount the contribution of each stream to the final result by properly taking its relative measurement error into account; however they do not need to satisfy a sum-to-one constraint $\sum_{s=1}^{S} w_{s,c,m_0} = 1$, as is conventionally considered by other authors.

This is an appealing result. Our framework unveils the probabilistic assumptions under stream weight-based formulations; furthermore, Eq. (7) provides a rigorous mechanism to select for each new measurement y_s and uncertainty estimate $(\mu_{e,s}, \Sigma_{e,s})$ all involved stream weights *fully adaptively*, *i.e.* with respect to both class label c and mixture component m.

3. Audio-Visual Speech Recognition

To demonstrate the applicability of the proposed fusion scheme we sucessfully apply it in Audio-Visual Automatic Speech Recognition (AV-ASR), a practical problem for which proper information fusion is very important.

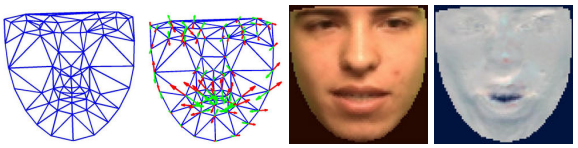

Figure 2: *Left*: Mean shape s_0 and the first eigenshape s_1. *Right*: Mean texture A_0 and the first eigentexture A_1.

3.1. Visual Front-end

Salient visual speech information can be obtained from the shape and the texture (intensity/color) of the speaker's visible articulators, mainly the lips and the jaw, which constitute the *Region Of Interest* (ROI) around the mouth [1].

We use *Active Appearance Models* (AAM) [13] of faces to accurately track the speaker's face and extract visual speech features from it, capturing both the shape and the texture of the face. AAM, which were first used for AV-ASR in [14], are generative models of object appearance and have been proven particularly effective in modeling human faces for diverse applications, such as face recognition or tracking. In the AAM scheme an object's shape is modeled as a wireframe mask defined by a set of landmark points $\{x_i, i = 1 \dots N\}$, whose coordinates constitute a shape vector s of length $2N$. We allow for deviations from the mean shape s_0 by letting s lie in a linear n-dimensional subspace, yielding $s = s_0 + \sum_{i=1}^{n} p_i s_i$. The deformation of the shape s to the mean shape s_0 defines a mapping $W(x; p)$, which brings the face exemplar on the current frame I into registration with the mean face template. After factoring out shape deformation, the face color texture registered with the mean face can be modeled as a weighted sum of "eigenfaces" $\{A_i\}$ as: $I(W(x; p)) \approx A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$, where A_0 is the mean texture of faces. Both the eigenshapes and their eigentextures are learned during a training phase.The first few of them extracted by such a procedure are depicted in Fig. 2.

Given a trained AAM, model fitting amounts to finding for each video frame I_t the parameters $\tilde{p}_t \equiv \{p_t, \lambda_t\}$ which minimize the squared texture reconstruction error $I_t(W(p_t)) - A_0 - \sum_{i=1}^{m} \lambda_{t,i} A_i$; efficient iterative algorithms for this non-linear least squares problem can be found in [13]. The fitting procedure employs a face detector to get an initial shape estimate for the first frame. To extract information mostly related to visual speech, we utilize a hierarchy of two AAM. The first *ROI-AAM* spans only the area around the mouth and is used to analyze in detail the ROI's shape and texture; however, the ROI-AAM covers too small an area to allow for reliable tracking. To pinpoint the ROI-AAM we use a second *Face-AAM* which spans the whole face and can reliably track the speaker in long video sequences. As visual feature vector for speech recognition we use the parameters $\tilde{p}_t$ of the ROI-AAM. We currently employ as uncertainty in the visual features the uncertainty in estimating the parameters of the corresponding non-linear least squares problem [15]; a better estimate can be obtained by using an extended Kalman filter-based visual tracker [16]. However, since AAM fitting is a non-linear optimization task, these methods tend to under-estimate the tracking error in case the AAM instantaneously mistracks the face and therefore heavy-tailed distributions might model the tracking error more reliably; we defer further study of this issue for future work.

3.2. Audio Front-end

In our expreriments, we use Mel Frequency Cepstral Coefficients (MFCC) and their time derivatives to represent the audio stream.

SNR	A	V	AV	P-AV	AV-UC	P-AV-UC
clean	100.0	68.7	95.1	95.4	97.0	99.6
10 dB	92.8	-	88.3	90.6	90.2	92.5
5 dB	73.9	-	84.5	87.2	86.8	89.1
0 dB	54.7	-	79.6	83.8	81.1	82.6

Table 1: Word Percent Accuracy (%) of classification experiments on CUAVE database for babble noise; Audio (A), Visual (V), baseline Audio-Visual (AV), Product HMM (P-AV), the proposed Audio-Visual Fusion with Uncertainty Compensation using multi-stream HMMs (AV-UC) and Product HMMs (P-AV-UC) .

Uncertainty is caused by the artificial addition of noise to the speech waveforms. An enhancement process provides us estimates of the clean features, namely those that would have been extracted from the clean waveforms. These are the features we base our classification decision on. In our framework, we consider the error of the enhancement process to be Gaussian. We have shown in Sec. 2 that if we can have an estimate of this error/uncertainty along with the clean feature estimate and use both in classification then audiovisual fusion could benefit. There are various enhancement techniques that can be applied for this purpose [7, 8, 2]. In our preliminary experiments, we regard uncertainty in our feature estimates to be zero-mean. To estimate its variance for each feature, we utilize the squared difference between the clean feature estimate and the clean feature which is also available in our case.

3.3. Audio-Visual Speech Recognition Experiments

The fusion approach proposed above is evaluated on the CUAVE audiovisual database [17] in the digit classification task. By contaminating the clean audio signal with babble noise from the NOISEX dataset [18] we extended the database including its noisy version. The audio feature vector included 13 static MFCC and their derivatives. Cepstral Mean Normalization has been applied to the static features. As far as the visual front-end is concerned, we form a visual feature vector by concatenating 6 shape and 12 texture features. For the acoustic and visual modeling of the observations we constructed 8-state left-right word HMM with a single mutlidimensional Gaussian as observation probability distribution per stream at each state. The models were trained on clean data. Feature uncertainty was taken into account during the testing phase as described by Eq. (5).

Classification results are shown in Table 1 for 0, 5 and 10dB SNR. Compensation for uncertainty (AV-UC, Audio-Visual Uncertainty Compensation) shows superior performance in noise compared to the audiovisual classification result (AV) obtained using the typical multistream HMMs. For variance compensation in the clean case, we only account for the visual observation error. Better estimates of the uncertainty in the visual features could in general lead to improved results. Further, to account for audio-visual speech asynchronicity we present experiments that utilize our scheme in conjuction with Product HMMs (P-AV and P-AV-UC). The conducted experiments show superior improvement compared to the synchronized case.

4. Perspective

The paper has shown that taking the feature uncertainty into account constitutes a fruitful framework for multimodal feature analysis tasks. This is especially true in the case of multiple complementary information streams, where having a good estimate of each stream's uncertainty at a particular moment allows for fully adaptive stream integration schemes, greatly facilitating information fusion. In order this approach to reach its full potential, reliable methods for dynamically estimating the feature observation uncertainty are needed. Ideally, the methods that we employ to extract features in pattern recognition tasks should accompany feature estimates with their respective errorbars. Although various authors have done progress in the area, much remains to be done before we fully understand the quantitative behavior under diverse environmental conditions of popular features commonly used in speech recognition.

5. Acknowledgments

We thank A. Potamianos for discussions and for providing the initial experimental setup for AV-ASR, I. Kokkinos for visual front-end discussions, K. Murphy for using his HMM toolkit, and J.N. Gowdy for the use of the CUAVE database.

6. References

[1] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[2] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, pp. 25–40, 2001.

[3] A. Potamianos, E. Sanchez-Soto, and K. Daoudi, "Stream weight computation for multi-stream classifiers," in *Proc. ICASSP*, 2006.

[4] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. ICASSP*, 2001.

[5] V. Digalakis, J.R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE TSAP*, pp. 431–442, 1993.

[6] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE TSAP*, vol. 2, no. 2, pp. 245–257, 1994.

[7] N.B Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE TSAP*, vol. 10, no. 3, pp. 158–166, 2002.

[8] L. Deng, J. Dropo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE TSAP*, vol. 13, no. 3, pp. 412–421, 2005.

[9] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[10] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. ICASSP*, 2001.

[11] Q. Huo and C. Lee, "A bayesian predictive approach to robust speech recognition," *IEEE TSAP*, pp. 200–204, 2000.

[12] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1–15, 2002.

[13] T.F. Cootes, G.J. Edwards, and Taylor C.J., "Active appearance models," *IEEE PAMI*, vol. 23, no. 6, pp. 681–685, 2001.

[14] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE PAMI*, vol. 24, no. 2, pp. 198–213, 2002.

[15] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes*, Cambridge Univ. Press, 1992.

[16] A. Blake and M. Isard, *Active contours : the application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*, Springer, 1998.

[17] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP*, 2002.

[18] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–252, 1993.