UTILITY OF AUXILIARY SENSOR DATA FOR SPEECH ENHANCEMENT

Sriram Srinivasan and Patrick Kechichian

Philips Research
Eindhoven, The Netherlands
{*sriram.srinivasan, patrick.kechichian*}*@philips.com*

ABSTRACT

In recent years, data from various auxiliary acoustic and non-acoustic sensors have been used for enhancing noisy speech. These include bone-conduction microphones, surface electromyographic sensors, ultrasonic imaging of facial movements, etc. The signal from such sensors is correlated with the speech signal to varying degrees, and unlike microphone data, is typically not affected by acoustic background noise, making its use attractive for speech enhancement. In this paper, we discuss the measurement of the utility of such data from an information-theoretic perspective, and quantify the information that is shared between clean speech and the auxiliary signal, which is not present in the observed noisy speech signal. The measure is applied to simultaneously recorded air- and bone-conducted speech data.

Index Terms— Speech enhancement, mutual information, bone conduction.

1. INTRODUCTION

Speech enhancement based on conventional single and multi-microphone techniques fails to deliver a sufficient amount of noise reduction under several realistic conditions, e.g., in reverberant environments, and under high levels of non-stationary background noise. The use of novel sensing modalities other than microphones to aid in the capture of speech in such adverse conditions has grown in recent years. Examples include surface electromyographic sensors that capture facial movements [1], ultrasonic sensors to capture tongue or lip movements [2, 3], non-audible murmur microphones to capture body-conducted sound [4], and bone-conduction microphones [5]. As these sensors capture some form of body movement related to speech production, their data is correlated to the clean speech signal, and more importantly, they are typically free from background noise. Another application where some of these sensors find use is the so-called silent speech interface, which enables speech communication even in the absence of any audible signal, e.g., for privacy or security, or due to disability.

The benefit provided by auxiliary sensors is generally evaluated in relation to the application in which they are employed, e.g., in terms of the improvement in the signal-to-noise ratio (SNR) when used for speech communication in noisy environments [5], or in terms of the recognition accuracy when applied to speech recognition systems [3, 6]. It is also of interest, however, to focus on how much information is intrinsically shared between the clean speech signal and the auxiliary sensor data, independent of the particular manner or application in which it is being used, e.g., to compare multiple auxiliary sensors. In this paper, we propose to use a more general information-theoretic measure to achieve this goal.

In prior work, mutual information (MI) has been used as a measure to study the relation between different sets of speech data. In [7, 8], the MI between the low and high band frequencies of speech was studied, e.g., to provide upper bounds on the performance of artificial bandwidth extension schemes operating without side-information. In [9], the estimated MI between clean and enhanced speech was used as a measure of speech intelligibility.

In contrast to these methods that measure the information shared between two signals, we are interested in quantifying the information shared between two signals that is not present in a third signal. For example, knowing how much information is shared between clean speech and the auxiliary sensor data, which is not present in the observed noisy data allows us to quantify the utility of the auxiliary sensor. We define two measures based on conditional MI, one that captures the utility of the auxiliary sensor alone, and another that captures the total system utility. We use these measures to analyze the utility of bone-conducted (BC) speech in the enhancement of noisy speech at different SNRs and for different locations of the BC sensor.

2. MEASURING AUXILIARY DATA UTILITY

Let X be a continuous random variable with probability density function (pdf) $p(x)$. The differential entropy of X is defined as [10, ch. 9]

$$h(X) = -\int p(x) \log p(x) dx. \tag{1}$$

The MI between two random variables X and Y with joint pdf $p(x, y)$ is given by

$$\begin{aligned} I(X,Y) &= \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx\, dy \\ &= h(X) + h(Y) - h(X,Y). \end{aligned} \tag{2}$$

$I(X, Y)$ quantifies the amount of information obtained about X when Y is observed, or vice-versa. $I(X, Y) \geq 0$ with equality when X and Y are uncorrelated. Given a third random variable Z (e.g., BC speech), we wish to determine the amount of information shared between X (e.g., clean speech) and Z, that is not present in Y (e.g., noisy speech). This would enable us to quantify the utility of Z in the estimation of X given a noisy observation Y. We denote this quantity as $U_X(Z|Y)$ and it is given by the conditional MI between X and Z given Y

$$\begin{aligned} U_X(Z|Y) &\equiv I(X,Z|Y) \\ &= h(X,Y) + h(Z,Y) - h(Y) - h(X,Y,Z). \end{aligned} \tag{3}$$

It is readily seen that $U_X(Z|Y)$ is non-negative. $U_X(Z|Y)$ can be interpreted as the additional information that Z contains about X, which is not available in Y. In a total system, both Z and Y will be employed to estimate X, which motivates the definition of a total utility measure for the estimation of X as

$$\begin{aligned} U_X(Z,Y) &\equiv I(X,Z|Y) + I(X,Y) \\ &= U_X(Z|Y) + I(X,Y), \end{aligned} \tag{4}$$

which quantifies the total unique information that Z and Y contain about X. Using (2) and (3), (4) can be rewritten as

$$U_X(Z,Y) = h(X) - h(X|Y,Z), \tag{5}$$

i.e., the total system utility $U_X(Z, Y)$ is the reduction in the uncertainty about X when Y and Z are observed. We consider a simple illustrative example to study the sensor and total system utility in our context. Let

$$\begin{aligned} &\mathbf{X} = (X_1, X_2),\ X_i \sim \mathcal{N}(0, \sigma_{x_i}^2),\ i = 1, 2, \\ &\mathbf{W} = (W_1, W_2),\ W_i \sim \mathcal{N}(0, \sigma_{w_i}^2),\ i = 1, 2, \\ &\mathbf{Y} = \mathbf{X} + \mathbf{W},\ \text{and} \\ &Z = X_1 + W_z,\ W_z \sim \mathcal{N}(0, \sigma_{w_z}^2). \end{aligned} \tag{6}$$

We assume X_1, X_2, W_1, W_2, and W_z to be mutually independent and that $\sigma_{w_z}^2 \ll \min_{i=1,2} \sigma_{w_i}^2$. In this example, $\mathbf{Y}$ is a noisy observation of $\mathbf{X}$, and Z contains partial information about $\mathbf{X}$, and in particular, more information about X_1 than $\mathbf{Y}$ but contains no information about X_2. For large values of $\sigma_{w_i}^2$, Z can contain more information about $\mathbf{X}$ than is present in $\mathbf{Y}$ as $\sigma_{w_z}^2 \ll \sigma_{w_i}^2$. The role of Z in this example is similar to that of BC speech, which is free from acoustic background noise, but only contains a partial description of the clean AC speech.

The vector $(\mathbf{X}, \mathbf{Y}, Z)$ is jointly Gaussian with zero mean and covariance matrix given by

$$\Sigma_{\mathbf{XY}Z} = \begin{pmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}Z} \\ \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{Y}} & \Sigma_{\mathbf{X}Z} \\ \Sigma_{\mathbf{X}Z}^T & \Sigma_{\mathbf{X}Z}^T & \Sigma_Z \end{pmatrix}, \tag{7}$$

where

$$\begin{aligned} \Sigma_{\mathbf{X}} &= \begin{pmatrix} \sigma_{x_1}^2 & 0 \\ 0 & \sigma_{x_2}^2 \end{pmatrix}, \\ \Sigma_{\mathbf{Y}} &= \begin{pmatrix} \sigma_{x_1}^2 + \sigma_{w_1}^2 & 0 \\ 0 & \sigma_{x_2}^2 + \sigma_{w_2}^2 \end{pmatrix}, \\ \Sigma_{\mathbf{X}Z} &= (\sigma_{x_1}^2 + \sigma_{w_z}^2, 0)^T,\ \text{and} \\ \Sigma_Z &= \sigma_{x_1}^2 + \sigma_{w_z}^2. \end{aligned} \tag{8}$$

The differential entropy of a d-dimensional Gaussian random variable with zero mean and covariance Σ is given by [10, ch. 9]

$$h(X) = \frac{1}{2} \log (2\pi e)^d |\Sigma|. \tag{9}$$

Using (2), (3), and (9), we have

$$\begin{aligned} I(\mathbf{X}, Z) &= \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}} \Sigma_Z|}{|\Sigma_{\mathbf{X}Z}|}, \\ I(\mathbf{X}, \mathbf{Y}) &= \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}} \Sigma_{\mathbf{Y}}|}{|\Sigma_{\mathbf{XY}}|},\ \text{and} \\ I(\mathbf{X}, Z|\mathbf{Y}) &= \frac{1}{2} \log \frac{|\Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}Z}|}{|\Sigma_{\mathbf{Y}} \Sigma_{\mathbf{XY}Z}|}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \Sigma_{\mathbf{Y}Z} &= \begin{pmatrix} \Sigma_{\mathbf{Y}} & \sigma_{x_1}^2 \\ & 0 \\ \sigma_{x_1}^2 \quad 0 & \sigma_{x_1}^2 + \sigma_{w_z}^2 \end{pmatrix},\ \text{and} \\ \Sigma_{\mathbf{XY}} &= \begin{pmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}} \\ \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{Y}} \end{pmatrix}, \end{aligned} \tag{11}$$

from which $U_{\mathbf{X}}(Z|\mathbf{Y})$ and $U_{\mathbf{X}}(Z, \mathbf{Y})$ can be obtained.

Consider the following parameter settings: $\sigma_x^2 = \sigma_{x_1}^2 = \sigma_{x_2}^2 = 1$ and $\sigma_{w_z}^2 = 10^{-3}$. Let $\sigma_w^2 = \sigma_{w_1}^2 = \sigma_{w_2}^2$. Figure 1 plots $I(\mathbf{X}, Z)$, $I(\mathbf{X}, \mathbf{Y})$, $U_{\mathbf{X}}(Z|\mathbf{Y})$ and $U_{\mathbf{X}}(Z, \mathbf{Y})$ for different values of σ_w^2. When the SNR is poor, e.g., for $10 \log_{10} \sigma_x^2/\sigma_w^2 = -20$ dB, $\mathbf{X}$ and Z share around 3.5 nats of information (dotted line), none of which is present in $\mathbf{Y}$ as it is dominated by noise. For low SNR values, Z therefore has a high utility, and the sensor utility $U_{\mathbf{X}}(Z|\mathbf{Y})$ (dashed curve) coincides with the total system utility $U_{\mathbf{X}}(Z, \mathbf{Y})$ (dash-dot curve). The utility of Z decreases with increasing SNR and approaches zero as $\mathbf{Y}$ provides a better description of $\mathbf{X}$ than Z, and the total system utility coincides with $I(\mathbf{X}, \mathbf{Y})$. For

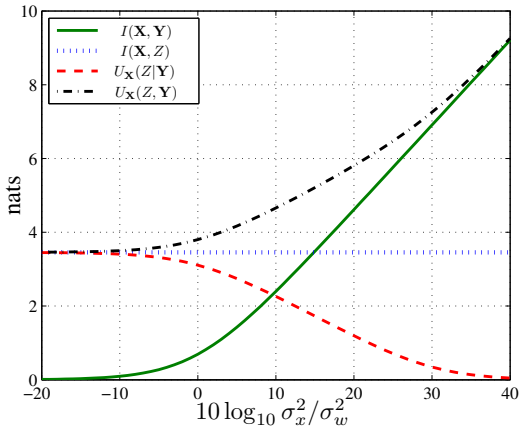

Fig. 1. $I(\mathbf{X},\mathbf{Y})$ (solid), $I(\mathbf{X},Z)$ (dotted), $U_{\mathbf{X}}(Z|\mathbf{Y})$ (dashed), and $U_{\mathbf{X}}(Z,\mathbf{Y})$ (dashed), in nats, for different SNRs.

a range of SNRs between approx. -10 dB and 30 dB in this example, both Z and $\mathbf{Y}$ contribute to the estimation of $\mathbf{X}$.

For speech enhancement using multi-modal sensors, $U_{\mathbf{X}}(Z|\mathbf{Y})$ can be used to evaluate the utility of the signal provided by each sensor, in relation to the noisy signal. In the next section, we consider simultaneously captured air-conducted (AC) and BC data, and compute the utility of BC data for two sensor locations.

3. UTILITY OF BONE-CONDUCTED DATA

3.1. Measurement Setup

Recordings were made in a dry room with a reverberation time of approximately 250 ms. A male and a female participant each wore BC sensors like the one shown in Fig. 2(b) positioned in the neck and temple regions shown in Fig. 2(a). A microphone placed 10 cm in front of the user's mouth captured the clean AC speech signal. Each participant was asked to utter sentences from the Harvard speech database [11] which were sequentially presented on a computer monitor for a duration of 6 min. The recorded signals were then downsampled to 8 kHz for further processing. To simulate noisy AC speech, white Gaussian noise with varying power was added to the recorded clean AC speech signal, producing an AC signal set with SNRs ranging from -40 to 40 dB in increments of 5 dB.

The sensor and total system utility measures for each position of the BC sensor over the simulated SNR range were estimated using the approach for conditional MI estimation in [12], which is based on the k-nearest neighbor (KNN) algorithm. This algorithm estimates the likelihood of a given sample based on the volume that encloses the k closest neighbors in the sample space. Here, the sample space consists of 12-dimensional Mel-Frequency Cepstral Coefficient (MFCC) feature vectors which are computed for overlapping segments of the 8-kHz AC and BC signals. The segment-length is 256 samples with an overlap of 50% and the value of k is set to the square-root of the resulting number of feature vectors, which corresponds to $k = 150$ in this case.

(a) Test location (b) Test prototype

Fig. 2. (a) Two locations considered in this paper: 1. Left side of neck on same transverse plane with the larynx and same frontal plane with shoulder; 2. Right squama of the temporal bone; (b) Prototype used during experiments.

3.2. Utility evaluation

Figure 3 plots the sensor and total system utility measures, $U_{\mathbf{X}}(\mathbf{Z}|\mathbf{Y})$ and $U_{\mathbf{X}}(\mathbf{Z},\mathbf{Y})$, of the BC sensor placed in the neck position, averaged over the male and female speakers, where the variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ correspond to clean AC, noisy AC and BC speech, respectively. For reference, the average values of $I(\mathbf{X},\mathbf{Z})$ and $I(\mathbf{X},\mathbf{Y})$ are also included. The curves show a similar behavior to those in Fig. 1. The MI $I(\mathbf{X},\mathbf{Z})$ curve serves as an upper bound to the maximum shared information between the BC and clean AC speech signal under ideal conditions when there is no noise leakage in the BC sensor. It also serves as a lower bound of the total system utility for low SNRs where noise dominates. Unlike $I(\mathbf{X},\mathbf{Z})$, the value of $I(\mathbf{X},\mathbf{Y})$ depends on the SNR, and can provide a quantitative measure of the amount of new information available as the SNR increases. For example, between -10 and 10 dB, 5 dB in SNR improvement translates to 0.25 more nats of information.

Examining the intersection points between the various curves in Fig. 3 also provides useful insights into the behavior of the system. For example, at an SNR of 1.8 dB where the $I(\mathbf{X},\mathbf{Y})$ and $U_{\mathbf{X}}(\mathbf{Z}|\mathbf{Y})$ curves intersect, the amount of information about the clean AC signal provided by the noisy AC signal equals that provided by the BC sensor signal. The point at which the $I(\mathbf{X},\mathbf{Y})$ and $I(\mathbf{X},\mathbf{Z})$ curves intersect indicates the SNR at which the amount of information provided by the noisy AC speech signal equals that provided by the BC sensor alone. For the neck region this corresponds to an SNR of almost 12 dB. Care has to be taken, however, with how this is interpreted since information *quantity* is being compared and not quality.

As the SNR improves, the utility of the BC sensor decreases as would be expected since the information it provides about the clean AC signal increasingly overlaps with that of the observed noisy AC speech signal. The total system utility $U_{\mathbf{X}}(\mathbf{Z}, \mathbf{Y})$ begins lower-bounded by $U_{\mathbf{X}}(\mathbf{Z}|\mathbf{Y})$ and starts increasing as the SNR improves since the amount of information about $\mathbf{X}$ provided by $\mathbf{Y}$ also increases. Eventually as the SNR improves further, the curves for $I(\mathbf{X}, \mathbf{Y})$ and $U_{\mathbf{X}}(\mathbf{Z}, \mathbf{Y})$ join when the utility of the BC sensor approaches zero.

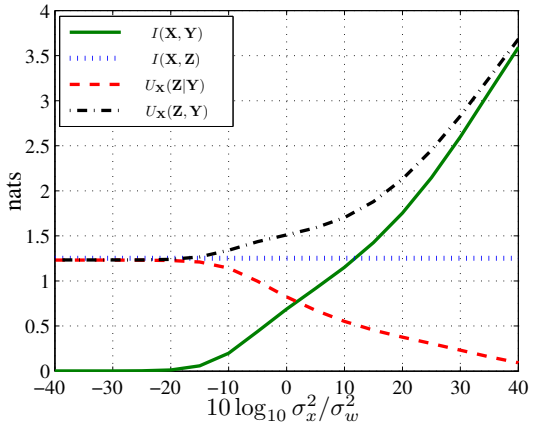

Fig. 3. $I(\mathbf{X}, \mathbf{Y})$ (solid), $I(\mathbf{X}, Z)$ (dotted), $U_{\mathbf{X}}(Z|\mathbf{Y})$ (dashed), and $U_{\mathbf{X}}(Z, \mathbf{Y})$ (dashed), in nats, for different SNRs.

To appreciate the utility measure as a tool for evaluating sensor placement, Fig. 4 compares the utility of bone-conducted speech captured at the neck and temple regions. The plot shows that the utility of the temple BC sensor is higher than that of the neck-positioned sensor. This is related to the fact that BC signals captured at the temple contain more information about the vocal tract shape. In contrast, a BC sensor placed at the neck primarily captures the vocal cord vibrations and weaker reflections from the vocal tract and hence contains less information about the clean AC speech signal [13].

4. CONCLUSION

This paper has presented an information-theoretic utility measure for auxiliary sensors in the context of speech capture in the presence of noise. This measure quantifies the amount of information provided by an auxiliary sensor about a clean speech signal which is missing from its noise-corrupted version. Experiments were performed using bone-conducted speech at two positions on the user's body and for different signal-to-noise ratios to highlight the benefits of the measure. It was observed that bone-conducted speech captured at the temple region had a slightly higher utility than that captured at the neck position. Future applications include the evaluation of different sensors and their placement in multi-modal systems for different speech enhancement applications.

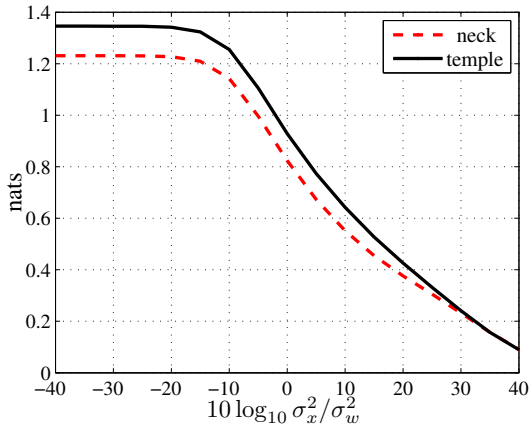

Fig. 4. $U_{\mathbf{X}}(\mathbf{Z}|\mathbf{Y})$ in nats, for neck (dashed), and temple (solid) regions.

5. REFERENCES

[1] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Communication*, vol. 52, no. 4, pp. 354–366, Apr. 2010.

[2] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, Apr. 2010.

[3] S. Srinivasan, B. Raj, and T. Ezzat, "Ultrasonic sensing for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Mar. 2010, pp. 5102 –5105.

[4] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Communication*, vol. 52, no. 4, pp. 301–313, Apr. 2010.

[5] P. Kechichian and S. Srinivasan, "Model-based speech enhancement using a bone-conducted signal," *J. Acoust. Soc. Amer.*, vol. 131, no. 3, 2012.

[6] J. Wang, A. Samal, J.R. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Mar. 2012, pp. 4985 –4988.

[7] M. Nilsson, S.V. Andersen, and W.B. Kleijn, "On the mutual information between frequency bands in

speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, June 2000, vol. 3, pp. 1327–1330.

[8] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, May 2002, vol. 1, pp. 237–240.

[9] J. Taghia, R. Martin, and R.C. Hendriks, "On mutual information as a measure of speech intelligibility," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Mar. 2012, pp. 65 –68.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, 1991.

[11] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended pratice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, Sept. 1969.

[12] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Phys. Rev. Lett.*, vol. 99, Nov. 2007.

[13] P. Tran, T. Letowski, and M. McBride, "Bone conduction microphone: Head sensitivity mapping for speech intelligibility and sound quality," in *Int. Conf. Audio, Language and Image Processing*, July 2008, pp. 107 – 111.