EFFECTIVE SPEECH/PAUSE DISCRIMINATION USING AN INTEGRATED BISPECTRUM LIKELIHOOD RATIO TEST

J. M. Górriz, J. Ramírez, J. C. Segura, C.G. Puntonet and L. García

Dept. of Signal Theory, Networking and Communications
University of Granada, Spain

ABSTRACT

This paper shows an effective voice activity detector based on a statistical likelihood ratio test defined on the integrated bispectrum of the signal. It inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: *i*) its computation as a cross spectrum leads to significant computational savings, and *ii*) the variance of the estimator is of the same order as that of the power spectrum estimator. The proposed method incorporates contextual information to the decision rule, a strategy that has reported significant improvements in speech detection accuracy and robust speech recognition applications. The experimental analysis conducted on the well-known AURORA databases has reported significant improvements over standardized techniques such as ITU G.729, AMR1, AMR2 and ESTI AFE VADs, as well as over recently published VADs.

1. INTRODUCTION

Currently, there are technology barriers inhibiting speech processing systems that work in extremely noisy conditions from meeting the demands of modern applications. These systems often require a noise reduction system in combination with a precise voice activity detector (VAD). Most of the algorithms for detecting presence of speech in a noisy signal only exploit the power spectral content of the signals and require knowledge of the noise power spectral density [1, 2, 3, 4]. One of the most important disadvantages of these approaches is that no *a priori* information about the statistical properties of the signals is used. Higher order statistics methods rely on an *a priori* knowledge of the input processes and have been considered for VAD since they can distinguish between Gaussian signals (which has a vanishing bispectrum) from non-Gaussian signals. However, the main limitations of bispectrum-based techniques are that they are computationally expensive and the variance of the bispectrum estimators is much higher than that of power spectral estimators for identical data record size. These problems were addressed by Tugnait [5, 6] who showed a computationally efficient and reduced variance statistical test based on the integrated polyspectra for detecting an unknown random, stationary, non-Gaussian signal in Gaussian noise. This paper advances in the field and shows an effective VAD based on a likelihood ratio test (LRT) that is defined on the integrated bispectrum of the noisy speech. The proposed approach also incorporates contextual information to the decision rule, a strategy first proposed in [7] that has reported significant benefits [8] and particularly, in robust speech recognition applications [9, 10, 11].

This work has been funded by the European Commission (HIWIRE, IST No. 507943) and the Spanish MEC project TEC2004-03829/FEDER.

2. INTEGRATED BISPECTRUM

The bispectrum of a discrete-time signal $x(t)$ is defined as:

$$B_x(\omega_1,\omega_2)=\sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}C_{3x}(i,k)\exp\{-j(\omega_1 i+\omega_2 k)\} \quad (1)$$

where $C_{3x}(i,k)=E\{x^*(t)x(t+i)x(t+k)\}$ is the third-order cumulant of the process $x(t)$. Note that, from the above definition, the third-order cumulant can be expressed as:

$$C_{3x}(i,k)=\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}B_x(\omega_1,\omega_2)\exp\{j(\omega_1 i+\omega_2 k)\}d\omega_1 d\omega_2 \quad (2)$$

Although the bispectra have all the advantages of cumulants/polyspectra, their direct use has two serious limitations: *i*) the computation of bispectra in the whole triangular region is huge, and *ii*) the two-dimensional (2-D) template matching score in the classification is impractical. To use efficiently bispectra, integrated bispectrum methods [5, 6] were proposed for different applications [12, 13].

2.1. Definition

Let $x(t)$ be a zero mean stationary random process. If we define $\tilde{y}(t)=x^2(t)-E\{x^2(t)\}$, the cross correlation between $\tilde{y}(t)$ and $x(t)$ is defined to be:

$$r_{\tilde{y}x}(k)=E\{\tilde{y}(t)x(t+k)\}=E\{x^2(t)x(t+k)\}=C_{3x}(0,k) \quad (3)$$

and its cross spectrum is given by:

$$S_{\tilde{y}x}(\omega)=\sum_{-\infty}^{\infty}C_{3x}(0,k)\exp\{-j\omega k\} \quad (4)$$

with

$$C_{3x}(0,k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}S_{\tilde{y}x}(\omega)\exp\{j(\omega_k)\}d\omega \quad (5)$$

If equations 2 and 5 are compared we obtain:

$$S_{\tilde{y}x}(\omega)=\frac{1}{2\pi}\int_{-\pi}^{\pi}B_x(\omega,\omega_2)d\omega_2=\frac{1}{2\pi}\int_{-\pi}^{\pi}B_x(\omega_1,\omega)d\omega_1 \quad (6)$$

Thus, the integrated bispectrum is defined as a cross spectrum between the signal and its square, and therefore, it is a function of a single frequency variable. It is easy to see that the bispectrum of a Gaussian process is identically zero, its integrated bispectrum is as

well. Hence, its computation as a cross spectrum leads to significant computational savings. But more important is that the variance of the estimator is of the same order as that of the power spectrum estimator [5].

2.2. Estimation

Let $\hat{S}_{yx}(\omega)$ denote a consistent estimator of $S_{yx}(\omega)$, where $y(t) = x^2(t) - E\{x^2(t)\}$. Given a finite data set $x(1), x(2), ..., x(N)$, the integrated bispectrum is normally estimated by dividing the sample sequence into segments or blocks [14]. Thus, the data set is divided into K_B non-overlapping segments each of size N_B samples so that $N = K_B N_B$. Then, the cross periodogram of the *i*th block of data is given by

$$\hat{S}_{yx}^{(i)}(\omega) = \frac{1}{N_B} X^{(i)}(\omega) \left[Y^{(i)}(\omega)\right]^* \tag{7}$$

where $X^{(i)}(\omega)$ and $Y^{(i)}(\omega)$ denote the discrete Fourier transform (DFT) of the *i*th block. Finally, the estimate is obtained by averaging K_B blocks

$$\hat{S}_{yx}(\omega) = \frac{1}{K_B} \sum_{i=1}^{K_B} \hat{S}_{yx}^{(i)}(\omega) \tag{8}$$

3. VOICE ACTIVITY DETECTION

This section addresses the VAD problem formulated in terms of a classical binary hypothesis testing framework:

$$\begin{array}{lll} H_0 & : & x(t) = n(t) \\ H_1 & : & x(t) = s(t) + n(t) \end{array} \tag{9}$$

In a two-hypothesis test, the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an observation vector $\hat{\mathbf{y}}$ to be classified, the problem is reduced to selecting the class (H_0 or H_1) with the largest posterior probability $P(H_i|\hat{\mathbf{y}})$. From the Bayes rule, a statistical LRT [1] is defined as:

$$L(\hat{\mathbf{y}}) = \frac{p_{\mathbf{y}|H_1}(\hat{\mathbf{y}}|H_1)}{p_{\mathbf{y}|H_0}(\hat{\mathbf{y}}|H_0)} \tag{10}$$

and the observation vector $\hat{\mathbf{y}}$ is classified as H_1 if $L(\hat{\mathbf{y}})$ is greater than $P(H_0)/P(H_1)$ otherwise it is classified as H_0.

Assuming the integrated bispectrum $\{S_{yx}(\omega) : \omega\}$ as the feature vector $\hat{\mathbf{y}}$ and to be independent zero-mean Gaussian variables in presence and absence of speech:

$$\begin{array}{l} p(S_{yx}(\omega)|H_0) = \frac{1}{\pi\lambda_0(\omega)} \exp\left[-\frac{|S_{yx}(\omega)|^2}{\lambda_0(\omega)}\right] \\ p(S_{yx}(\omega)|H_1) = \frac{1}{\pi\lambda_1(\omega)} \exp\left[-\frac{|S_{yx}(\omega)|^2}{\lambda_1(\omega)}\right] \end{array} \tag{11}$$

the evaluation of the test defined in 10 only requires to estimate the integrated bispectrum of the noisy signal and its variance. Thus, taking logarithms in 10 and substituting the model defined in 11 we obtain:

$$\begin{array}{c} \Phi(\hat{\mathbf{y}}) = \sum_{\omega} \log\left(\frac{p(S_{yx}(\omega)|H_1)}{p(S_{yx}(\omega)|H_0)}\right) = \\ \sum_{\omega} \left\{ \left(1 - \frac{\lambda_0(\omega)}{\lambda_1(\omega)}\right) \frac{|S_{yx}(\omega)|^2}{\lambda_0(\omega)} - \log\left(\frac{\lambda_1(\omega)}{\lambda_0(\omega)}\right) \right\} \end{array} \tag{12}$$

Finally, if we define the *a priori* and *a posteriori* variance ratios as:

$$\xi(\omega) = \frac{\lambda_1(\omega)}{\lambda_0(\omega)} - 1 \qquad \gamma(\omega) = \frac{|S_{yx}(\omega)|^2}{\lambda_0} \tag{13}$$

equation 12 can be expressed in a more compact form:

$$\begin{array}{c} \Phi(\hat{\mathbf{y}}) = \sum_{\omega} \left[\left(1 - \frac{1}{1+\xi(\omega)}\right) \gamma(\omega) - \log\left(1 + \xi(\omega)\right) \right] = \\ = \sum_{\omega} \left[\frac{\xi(\omega)\gamma(\omega)}{1+\xi(\omega)} - \log\left(1 + \xi(\omega)\right) \right] \end{array} \tag{14}$$

Thus, the two key issues to evaluate the proposed LRT are: *i*) the estimation of the integrated bispectrum by means of a finite data set, and *ii*) the computation of the variances $\lambda_0(\omega)$ and $\lambda_1(\omega)$ of the integrated bispectrum under H_0 and H_1 hypothesis.

4. VARIANCE OF THE INTEGRATED BISPECTRUM

The properties of the bispectrum estimators has been discussed in [14, 15]. The test proposed in the previous section and the model assumed in equation 11 are justified since for large N_B, the estimate $S_{yx}^{(i)}(\omega_m)$ is complex Gaussian and independent of $S_{yx}^{(i)}(\omega_n)$ for $m \neq n$ $(m, n = 1, 2, ..., N_B/2 - 1)$. Moreover, its mean and variance for large values of N_B and K_B can be approximated [5] by:

$$\begin{array}{l} E\left\{\hat{S}_{yx}(\omega)\right\} \approx S_{yx}(\omega) \\ \text{var}\left\{\Re\left[\hat{S}_{yx}^{(i)}(\omega)\right]\right\} \approx \frac{1}{2K_B}\left[S_{yy}(\omega)S_{xx}(\omega) + \Re\left\{S_{yx}^2(\omega)\right\}\right] \\ \text{var}\left\{\Im\left[\hat{S}_{yx}^{(i)}(\omega)\right]\right\} \approx \frac{1}{2K_B}\left[S_{yy}(\omega)S_{xx}(\omega) - \Re\left\{S_{yx}^2(\omega)\right\}\right] \end{array} \tag{15}$$

and the estimation of $\lambda_0(\omega)$ and $\lambda_1(\omega)$ requires to compute $S_{xx}(\omega)$ and $S_{yy}(\omega)$ under H_0 and H_1 hypothesis. It can be shown [5, 6] that:

$$\lambda_0(\omega) = \frac{1}{K_B}\left[2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4\delta(\omega)\right] S_{nn}(\omega) \tag{16}$$

$$\begin{array}{c} \lambda_1(\omega) = \frac{1}{K_B}\left[S_{ss}(\omega) + S_{nn}(\omega)\right] \\ \left[2S_{ss}(\omega) * S_{ss}(\omega) + 2S_{nn}(\omega) * S_{nn}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega)\right] \end{array} \tag{17}$$

Finally, a way to estimate the integrated bispectrum of the clean signal, $S_{ss}(\omega)$, is needed. In this paper, a method combining Wiener filtering and spectral subtraction is used to estimate $S_{ss}(\omega)$ in terms of the integrated bispectrum of the noisy signal $S_{xx}(\omega)$. During a short initialization period, the integrated bispectrum of the residual noise $S_{nn}(\omega)$ is estimated assuming a short non-speech period at the beginning of the utterance. Note that, $S_{nn}(\omega)$ can be computed in terms of the DFT of the noisy signal $x(t) = n(t)$. After the initialization period, the integrated bispectrum of the noisy signal $S_{xx}(\omega)$ is computed for each frame through equations 7 and 8 and $S_{ss}(\omega)$ is then obtained by applying a denoising process. Denoising consists of a previous smoothed spectral subtraction followed by Wiener filtering. Figure 1 shows a block diagram for the estimation of the denoised integrated bispectrum $S_{ss}(\omega)$ through the noisy signal $S_{xx}(\omega)$. It is worthwhile clarifying that $S_{nn}(\omega)$ is not only estimated during the initialization period but also updated during non-speech frames based on the VAD decision. Thus, the denoising process consists of the following stages:

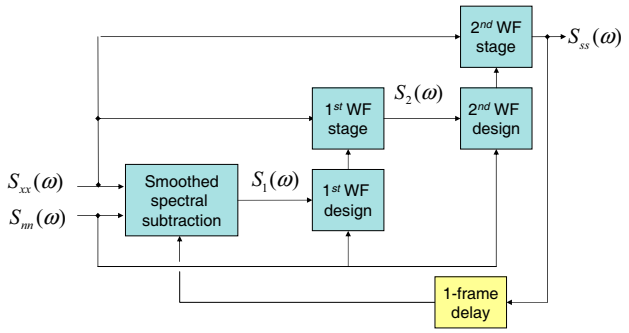

Fig. 1. Estimation of $S_{ss}(\omega)$ via smoothed spectral subtraction and Wiener filtering

1. Spectral subtraction.

$$S_1(\omega) = L_s S_{ss}(\omega) + (1 - L_s)\max(S_{xx}(\omega) - \alpha S_{nn}(\omega), \beta S_{xx}(\omega)) \tag{18}$$

2. First WF design and filtering.

$$\begin{aligned} \mu_1(\omega) &= S_1(\omega)/S_{nn}(\omega) \\ W_1(\omega) &= \mu_1(\omega)/(1+\mu_1(\omega)) \\ S_2(\omega) &= W_1(\omega) S_{xx}(\omega) \end{aligned} \tag{19}$$

3. Second WF design and filtering.

$$\begin{aligned} \mu_2(\omega) &= S_2(\omega)/S_{nn}(\omega) \\ W_2(\omega) &= \max(\mu_2(\omega)/(1+\mu_2(\omega)), \beta) \\ S_{ss}(\omega) &= W_2(\omega) S_{xx}(\omega) \end{aligned} \tag{20}$$

where $L_s = 0.99$, $\alpha = 1$ and $\beta = 10^{(-22/10)}$ is selected to ensure a -22dB maximum attenuation for the filter in order to reduce the high variance musical noise that normally appears due to rapid changes across adjacent frequency bins.

5. INTEGRATED BISPECTRUM LRT FOR VOICE ACTIVITY DETECTION

The proposed VAD is described as follows. The input signal $x(t)$ sampled at 8 kHz is divided into overlapping windows each of size $N = K_B N_B$ samples. A typical value of the window size in order to get precise estimations of the integrated bispectrum is about 0.2 seconds. The best tradeoff between block averaging (K_B) and spectral resolution (N_B) will be discussed in next sections.

Figure 2 illustrates the way the signal is processed and the block of data the decision is made for. Note that, the decision is made for a T-sample data block around the mid-point of the analysis window where T is the "frame-shift". Thus, a large data set is used to estimate the integrated bispectrum by averaging K_B successive blocks of data while the decision is made for a shorter data set. As in most of the standardized VADs [16, 17, 18] the frame-shift is 80 samples so that the VAD frame rate is 100 Hz.

After having estimated the integrated bispectrum $S_{ss}(\omega)$ of the clean signal, $\lambda_0(\omega)$ and $\lambda_1(\omega)$ are computed by evaluating the convolution operations required by equations 16 and 17. Then, the *a priori* and *a posteriori* variance ratios as defined in equation 13 can be estimated and the VAD decision rule is performed by comparing

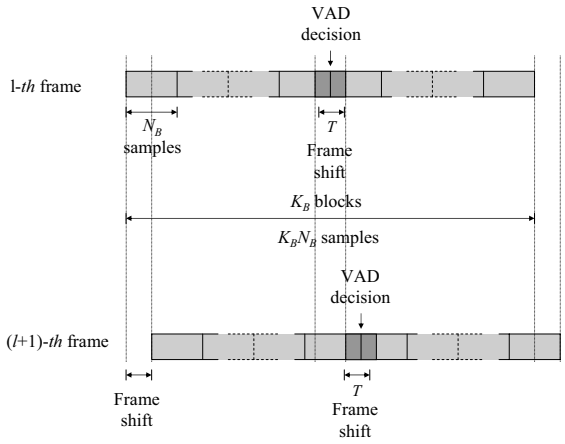

Fig. 2. Integrated bispectrum estimation by block averaging and VAD decision.

the LRT defined in equation 14 to a given threshold η. If the LRT is greater than the threshold η, the frame is classified as speech, otherwise it is classified as non-speech. Once the VAD decision is made for the frame being processed, the estimation of the integrated bispectrum of the noise is updated during non-speech periods in order to track non-stationary noisy environments:

$$S_{nn}(\omega) = L_n S_{nn}(\omega) + (1 - L_n) S_{xx}(\omega) \tag{21}$$

where L_n= 0.98.

6. EXPERIMENTAL ANALYSIS

The receiving operating characteristics (ROC) curves are frequently used to completely describe the VAD error rate. They show the tradeoff between speech and non-speech detection accuracy as the decision threshold varies [9]. The AURORA subset of the original Spanish SpeechDat-Car database [19] was used in this analysis. This database contains 4914 recordings using close-talking and distant microphones from more than 160 speakers. The files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions with average SNR values between 25dB and 5dB. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined in each noise condition being the actual speech frames and actual speech pauses determined by hand-labelling the database on the close-talking microphone.

Figure 3 shows the ROC curve of the proposed VAD and other frequently referred algorithms [1, 2, 3, 4] for recordings from the distant microphone in high noisy conditions. The working points of the G.729, AMR and AFE VADs are also included. Note that increasing the number of blocks (K_B) in the block averaging integrated bispectrum (BA-IBI) LRT VAD leads to a shift-up and to the left of the ROC curve in the ROC space. The proposed method shows clear improvements in detection accuracy over standardized VADs and over a representative set of recently published VAD algorithms [1, 2, 3, 4]. Among all the VAD examined, our VAD yields the lowest false alarm rate for a fixed non-speech hit rate and also, the highest non-speech hit rate for a given false alarm rate. The benefits are especially important over G.729, which is used along with

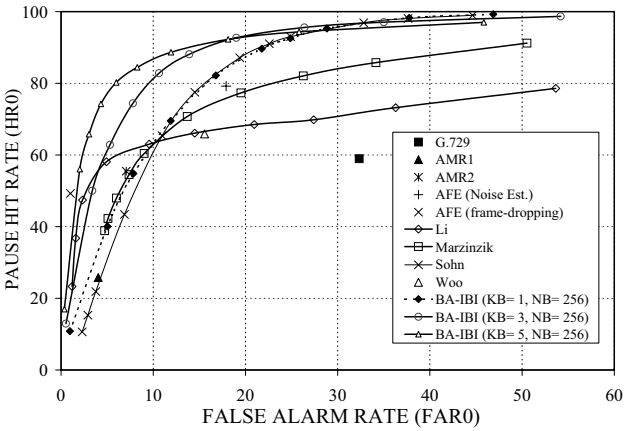

Fig. 3. ROC curves obtained in high noise conditions.

a speech codec for discontinuous transmission, and over Li's algorithm [3], that is based on an optimum linear filter for edge detection. The proposed VAD also improves Marzinzik's VAD [2] that tracks the power spectral envelopes, and the statistical VAD proposed by Sohn *et al.* [1], that formulates the decision rule by means of a statistical LRT defined on the power spectrum of the noisy signal.

7. CONCLUSIONS

This paper showed a voice activity detector for improving speech detection robustness in noisy environments. The proposed method is based on a statistical LRT defined on the integrated bispectrum of the signal which is defined as a cross spectrum between the signal and its square and inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: *i*) its computation as a cross spectrum leads to significant computational savings, and *ii*) the variance of the estimator is of the same order as that of the power spectrum estimator. It incorporates contextual information to the decision rule, a strategy that has reported significant improvements in speech detection accuracy and robust speech recognition applications. The experimental analysis conducted on the well-known AURORA databases has reported significant improvements over standardized techniques such as ITU G.729, AMR1, AMR2 and ESTI AFE VADs, as well as over recently published VADs.

8. REFERENCES

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3, 1999.

[2] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[3] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[4] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[5] J. K. Tugnait, "Detection of non-gaussian signals using integrated polyspectrum," *IEEE Trans. on Signal Processing*, vol. 42, no. 11, pp. 3137–3149, 1994.

[6] J. K. Tugnait, "Corrections to detection of non-gaussian signals using integrated polyspectrum," *IEEE Trans. on Signal Processing*, vol. 43, no. 11, pp. 2792–2793, 1995.

[7] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. of EUROSPEECH 2003*, Geneva, Switzerland, September 2003, pp. 3041–3044.

[8] A. Sangwan, W.P. Zhu, and M.O. Ahmad, "Improved voice activity detection via contextual information and noise suppression," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2005, pp. 868–871.

[9] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[10] J. Ramírez, José C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "An effective subband osf-based vad with noise reduction for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, 2005.

[11] J. Ramírez, José C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[12] X. Zhang, Y. Shi, and Z. Bao, "A new feature vector using selected bispectra for signal classification with application in radar target recognition," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1875–1885, 2001.

[13] X. Liao and Z. Bao, "Circularly integrated bispectra: Novel shift invariant features for high-resolution radar target recognition," *Electronics Letters*, vol. 34, no. 19, pp. 1879–1880, 1998.

[14] D. R. Brillinger and M. Rosenblatt, *Spectral Analysis of Time Series*, chapter Computation and interpretation of k-th order spectra, Wiley. New York, 1968.

[15] D.R. Brillinger, *Time series data analysis and theory*, New York: Holt, Rinehart and Winston, 1975.

[16] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[17] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[18] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2002.

[19] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proceedings of the II LREC Conference*, 2000.