GAUSSIAN POWER FLOW ORIENTATION COEFFICIENTS FOR NOISE-ROBUST SPEECH RECOGNITION

Branislav Gerazov and Zoran Ivanovski

Faculty of Electrical Engineering and Information Technologies,
Ss. Cyril and Methodius University Skopje, Macedonia

ABSTRACT

Spectro-temporal features have shown a great promise in respect to improving the noise-robustness of Automatic Speech Recognition (ASR) systems. The common approach uses a bank of 2D Gabor filters to process the speech signal spectrogram and generate the output feature vector. This approach suffers from generating a large number of coefficients, thus necessitating the use of feature dimensionality reduction. The proposed Gaussian Power flow Orientation Coefficients (GPOCs) use an alternative approach in which only the largest coefficients output from a bank of 2D Gaussian kernels are used to describe the spectro-temporal patterns of power flow in the auditory spectrogram. Whilst reducing the size of the feature vectors, the algorithm was shown to outperform traditional feature extraction methods, even a reference spectro-temporal approach, for low SNRs. Its performance for high SNRs is comparable but inferior to traditional ASR frontends, while falling behind state-of-the-art algorithms in all noise scenarios.

Index Terms— ASR, noise-robust, spectro-temporal, 2D Gaussian, kernel

1. INTRODUCTION

The Automatic Speech Recognition (ASR) systems of today are increasingly deployed on mobile platforms and plunged in to the noise filled world of busy streets, supermarkets, train stations, and cocktail parties. This "hostile" environment makes noise-robustness a critical parameter for ASR system design, fuelling an on-going research effort evident by the large amount of work done in this area [1]. One of the ways to address the noise-robustness of modern ASR systems is to develop features that are innately robust to noise. Significant developments towards this goal have been powered by modeling speech signal analysis in the human auditory system.

Traditional features such as the Mel-Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) have been based on modeling the peripheral auditory system. Further improvements in noise-robustness were made by modeling the intricate detail of nerve firing patterns, at the borderline between the peripheral and central auditory systems, by algorithms such as the classic Zero-Crossings with Peak Amplitudes (ZCPA), and the Subband Spectral Centroid Histogram (SSCH) [2].

In the last decade there has been an increasing shift in paradigm in ASR feature extraction towards modeling the central auditory system. Though the sensitivity of cortical neurons to spectral modulation components was long known and used in feature extraction; recent studies have additionally shown that groups of neurons are attuned to specific movements of energy in the speech spectrum, as described by the spectro-temporal receptive fields (STRF) model [3]. This has led to new approaches to feature extraction based on banks of 2D spectro-temporal Gabor filters that outperform traditional features in noise-robustness [4 – 6]. The common approach used is to pass the speech spectrogram through the filter bank and concatenate the filter outputs as feature vectors. This generates feature vectors with a large size, thus necessitating the use of feature dimensionality reduction strategies, such as feature selection, multi-layer perceptrons (MLP), and principle component analysis (PCA). For example, the state-of-the-art Gabor Tensor Cepstral Coefficients use a bank of 16 Gabor filters to generate 640 coefficients that are reduced to 26 using Nonnegative Tensor PCA with sparse constraints, the final size being 78 with the addition of delta and acceleration coefficients [6].

This paper proposes a new feature extraction algorithm that takes a novel approach to spectro-temporal analysis. The features, called the Gaussian Power flow Orientation Coefficients (GPOCs), are built around a set of 2D Gaussian kernels generated at a fixed scale with varying directionality. The novelty in the approach is that the algorithm uses only the directionality of the Gaussian kernel with the largest power output to describe the orientation of the power flow at each point in the spectrogram. This generates a novel view of the speech spectrogram termed the Power flow Orientation Spectrogram (POS) which efficiently captures both the movement of formants and the occurrence of plosives. The POS forms the basic set of GPOCs, and their number depends only on the number of bands in the auditory spectrogram, not the number of spectro-temporal kernels, eliminating the need of feature dimensionality reduction. The basic set is augmented with

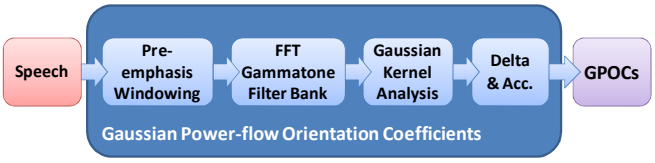

Fig. 1 – The structure of the Gaussian Power flow Orientation Coefficient extraction algorithm.

coefficients obtained at a larger time-scale, as well as delta and acceleration coefficients. The GPOC algorithm was shown to outperform traditional based feature extraction methods, for low SNRs, and has decreased performance compared to state-of-the-art ASR features.

2. GPOC ALGORITHM STRUCTURE

The general structure of the GPOC feature extraction algorithm is shown in Fig. 1. The auditory spectrogram is first calculated by the first two modules using well known methods from traditional ASR feature extraction schemes and a bank of 17 Gammatone filters. The filters' centre frequencies span the frequency range of 200 – 4000 Hz, and were calculated using the Glasberg and Moore Equivalent Rectangular Bandwidth (ERB) filter model, as implemented in [7]. The spectro-temporal patterns present in the auditory spectrogram are then analyzed with a set of Gaussian kernels and described through orientation coefficients for the power flow in each frequency band and frame index. The final module calculates and adds the dynamic coefficients.

3. GAUSSIAN KERNEL DESIGN

At the core of the proposed GPOCs are the Gaussian kernels derived from the 2D Gaussian function (1) as given in [8], where t_θ and f_θ are defined in (2) and (3). Here t and f are the time and frequency indexes, while σ is the standard deviation in respect to time σ_t. The standard deviation in respect to frequency σ_f is given implicitly by the aspect ratio r which is defined in (4) and describes the elipticity of the 2D Gaussian function. The rotation angle of the long axis of the elliptical 2D Gaussian is determined by θ. The Gaussian function was selected in favor to the commonly used Gabor function as it gave better experimental results.

$$G(t,f,\sigma,r,\theta) = \frac{1}{\sqrt{\pi r}\sigma} e^{-\frac{1}{2}\left(\frac{t_\theta^2}{\sigma^2} + \frac{f_\theta^2}{\sigma^2/r^2}\right)} \quad (1)$$

$$t_\theta = t \cdot \cos\theta + f \cdot \sin\theta \quad (2)$$

$$f_\theta = -t \cdot \sin\theta + f \cdot \cos\theta \quad (3)$$

$$r = \sigma_t / \sigma_f \quad (4)$$

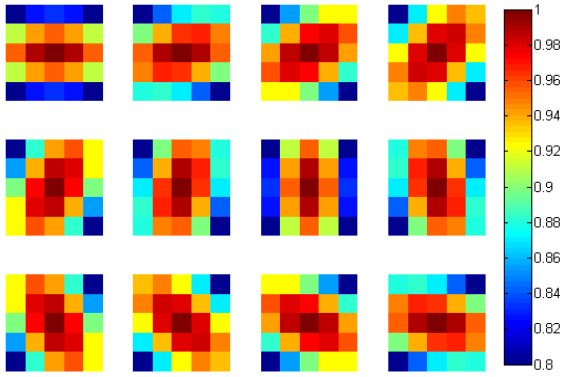

Fig. 2 – The set of Gaussian kernels used in the GPOC feature extraction algorithm.

The 2D Gaussian function G was used to generate the bank of Gaussian kernels k_G used in the GPOC algorithm according to (5), for constant σ and r, and a range of rotation angles θ_i at steps of θ_{step}, that cover the range 0^o – 180^o, (6). Because the Gaussian kernels are used to process the auditory spectrogram, here t_n and f_m refer to the frame number and the frequency band of the spectrogram, and N and M define the size of the kernel. The 5x5 k_G kernel bank with $\sigma = 9$, $r = 1.75$, and $\theta_{step} = 15^o$, used in GPOC is shown in Fig. 2.

$$k_G(t_n, f_m, \theta_i) = \begin{cases} G(t_n, f_m, \sigma, r, \theta_i) & \text{for } -N \le t_n \le N, -M \le f_m \le M \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$\theta_i = i \cdot \theta_{step} \qquad \text{for } i \in \{0,1,2,\ldots,\frac{180}{\theta_{step}} - 1 \quad (6)$$

4. POWER ORIENTATION CALCULATION

Each of the Gaussian kernels in the bank k_G is convolved with the auditory spectrogram as given in (7). Here * denotes convolution, J is a matrix of ones the size of the spectrogram, and the operator ./ denotes element-wise division of the matrixes. This normalizes the output coefficients, compensating for the influence of the zero-padding at the edges of the spectrogram for the different rotations of the kernels. This operation gives a set of output coefficient matrixes S_{oi} each with the same size as the spectrogram. The dominant power flow orientation θ_{max} for each time-frequency point (t, f) in the spectrogram is then set to the orientation θ_i of the kernel which generated the largest output coefficient at that point, (8).

$$S_{oi}(t,f,\theta_i) = S(t,f) * k_{Xi}(t_n, f_m, \theta_i) \,./\, J(t,f) * k_{Xi}(t_n, f_m, \theta_i) \quad \forall i \quad (7)$$

$$\theta_{\max}(t,f) = \arg\max_{\theta_i}\{S_{oi}(t,f,\theta_i)\} \quad (8)$$

The $\theta_{max}(t,f)$ for each point of the spectrogram gives the Power Flow Orientation Spectrogram (POS) which is a novel representation of the spectro-temporal content of the speech signal. An example POS of the utterance "makedonski" /mak□d□nski/ is shown in Fig. 3 juxtaposed to the calculated auditory spectrogram. The colorbar used to depict the POS wraps around giving points with more horizontally oriented power flow darker blue color, while more vertical flow patterns are presented with cyan. It can be seen that the horizontal power flows are closely connected to the movement of formants in the speech spectrogram, and they can be readily recognized in the POS. At the same time, the vertical power distributions of the plosive /k/ appearing twice in the utterance are also crisply captured.

The POS constitutes the basic set of GPOCs. To augment it, orientation coefficients are calculated at a larger scale by downsampling the auditory spectrogram with a factor of *d* in respect to time. This gives the scaled set of GPOCs which, together with the basic set, makes up the static set numbering a total of 34 coefficients. Figure 3 shows the auditory spectrogram and the basic and scaled sets of GPOCs. The static GPOCs demonstrate little correlation as can be seen from the Pearson correlation matrix given in Fig. 4. Indeed the use of decorrelation methods such as the Discrete Cosine Transform, yield a set of largely correlated coefficients, while PCA decreases ASR system performance.

Finally, dynamic GPOCs are calculated by estimating the first and second time derivatives of the static set of GPOCs using linear regression. The number of coefficients taken into account in estimating the first derivative was ±10 for the basic set and ±30 for the scaled set, to take into account the downsampling. For estimating the acceleration ±1 coefficients were used. Adding the delta and acceleration coefficients gives a vector length of 102.

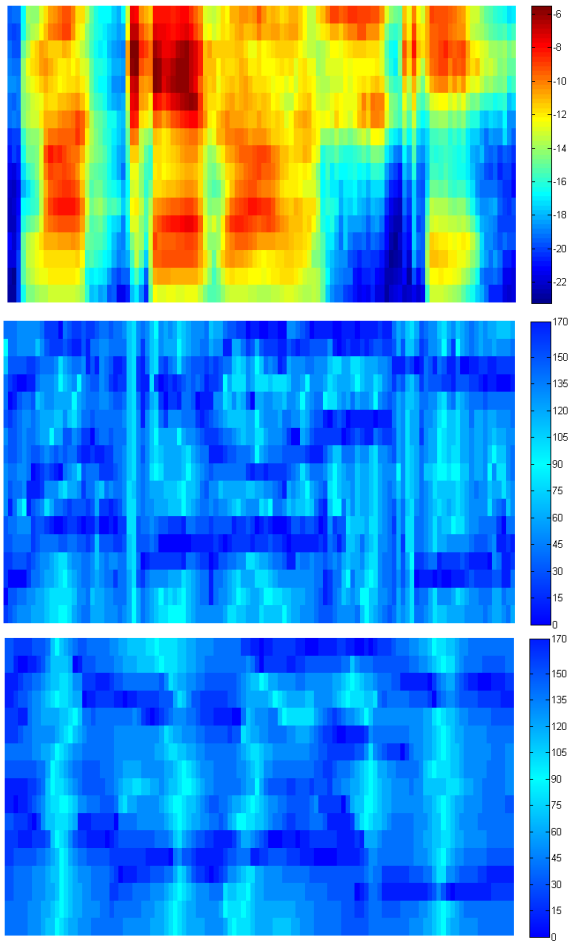

Fig. 3 – Example auditory spectrogram (*top*), power flow orientation spectrogram (*middle*) and its scaled version (*bottom*) for the utterance "makedonski" /mak□d□nski/.

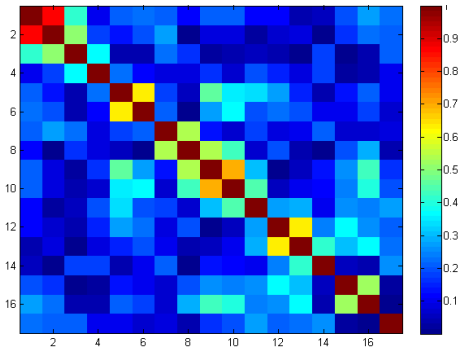

Fig. 4 – Pearson correlation matrix for the basic and whole set of GPOCs calculated for the utterance "makedonski" /mak□d□nski/.

5. EXPERIMENTAL SETUP

The Aurora 2 experimental framework was used to optimize and assess the performance of the proposed noise-robust GPOC feature extraction algorithm [9]. Aurora 2 is a connected digit recognition ASR task that uses the TIDigits database, which is noised using a selection of 8 different real-world noises over a range of signal to noise ratios (SNRs). Two training modes are defined in Aurora 2, one using clean recordings, the other multicondition (clean and noisy) training data. From the three test sets, set A is noised with the same four noises used in multicondition training, while set B uses four different noise types. Set C includes additional convolutional distortion. Aurora also includes a reference HMM recognition back-end. ASR system performance is evaluated using the Word Recognition Accuracy (WRA) as calculated by equation (9). Here *N* is the total number of words; *D* is the number of deletions; *S* is the number of substitutions, and *I* is the number of insertions.

$$Acc = \frac{N - D - S - I}{N} \cdot 100\% \qquad [\%] \quad (9)$$

6. REFERENCE FRONTENDS

Five well-known frontends were used as reference: MFCCs, Power Normalized Cepstral Coefficients (PNCCs), Subband Spectral Centroid Histograms (SSCHs), Gabor filter bank features (GBFB), and ETSI's Advanced Frontend (AFE). MFCC results were generated with their implementation included with Aurora. PNCCs are state-of-the-art noise-robust features based on MFCCs; they incorporate peak power normalization and medium-duration power bias subtraction with power flooring [10]. SSCHs are based on dominant-frequency information extracted from the locations of the subband spectral centroids and the power around them, [2]. The Gabor filter bank features (GBFB) represent the classic spectro-temporal approach, [5]. They use a bank of 41 Gabor filters to process a 23-channel Mel filter auditory spectrogram, generating 943 output coefficients which are reduced to 311 by subsampling. Finally, the state-of-the art AFE uses an augmented MFCC scheme, featuring blocks for Wiener noise reduction and blind equalization, [11].

7. RESULTS

The Aurora 2 framework was first used to optimize the parameters of the GPOC algorithm. In this procedure, the size of the kernels, the rotation step size θ_{step}, the σ and r of the 2D Gaussian function, the downsampling factor d, as well as the bandwidth of the Gammatone filters were all varied in a chosen range to maximize WRA. The final chosen parameters used to generate the Gaussian kernel bank k_G as shown in Fig. 2, were: kernel size – 5x5, $\theta_{step} = 15^o$, $\sigma = 9$, $r = 1.75$, $d = 3$, ERB scaling factor $BW = 0.75$.

The optimized GPOC was compared to the reference frontends. The results of this comparison are shown with an overlapped column chart in Fig. 5, in respect to the clean, multicondition and global average WRAs. The figure shows ETSI's AFE boasting the best performance, followed by PNCC. The GPOC algorithm is next in respect to global average and clean training WRAs, with convincingly better results for clean training than MFCC, SSCH, and GBFB. Compared to MFCCs, the average improvement in WRA of the GPOC algorithm for clean training is 26%. In the multicondition training mode, because of reduced improvements in performance, relative to improvements in the reference algorithms, GPOCs fall behind overall.

A clearer picture of the performance of the GPOC algorithm can be obtained from Fig. 6, which shows plots of average WRAs in respect to SNR for the three test sets (A, B and C) in clean training mode, for the considered frontends. In the plots, GPOC performance can be seen to hover between that of AFE and PNCC, and GBFB, MFCC and SSCH, with AFE clearly dominating in all noise scenarios. More specifically, GPOCs generally outperform MFCC, GBFB and SSCH in SNRs bellow 15 dB. In multicondition training mode, not shown due to the limited

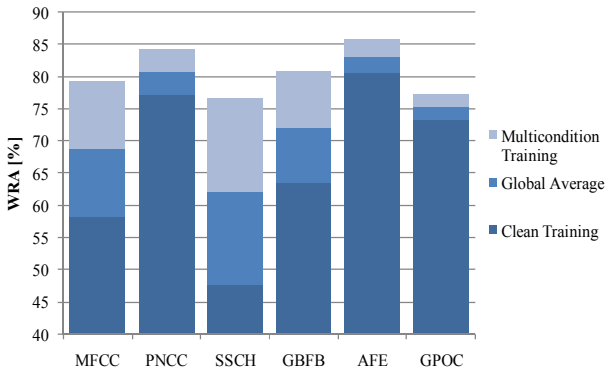

Fig. 5 – Comparison of the average WRA between the set of reference frontends and the GPOC algorithm.

space available, the WRA of all of the frontends can be seen to converge, with GPOC falling behind again for high SNRs.

Table 1 gives average WRAs for clean training mode in respect to SNR. Relative improvements of WRA of the GPOC algorithm compared to the reference frontends are given in Table 2. The cells showing GPOC outperforming the reference frontends are highlighted. Compared to MFCCs, the average improvement in accuracy of the GPOC algorithm for -5, 0, 5 and 10 dB is 156%, 191%, 94%, and 32%, with the highest improvement of 1215% for Babble noise in test set A. For an SNR of 20 dB and for clean speech, the GPOC algorithm falls short and delivers reduced performance compared to all reference frontends.

Table 1 – Average WRA of the GPOC algorithm and the reference frontends in clean training mode per SNR

	MFCC	PNCC	SSCH	GBFB	AFE	GPOC
AVG	58.27	77.25	64.69	63.48	80.67	73.43
-5 dB	8.53	22.59	11.45	9.34	30.32	21.80
0 dB	17.09	53.39	22.93	19.12	62.49	49.70
5 dB	38.61	79.44	48.15	47.74	84.57	74.76
10 dB	65.51	91.79	79.70	78.43	93.23	86.60
15 dB	85.04	96.28	93.65	92.86	96.66	91.59
20 dB	94.07	97.95	97.53	97.38	98.15	93.60
clean	99.03	99.28	99.39	99.50	99.23	95.97

Table 2 – Relative improvement of WRA in % of GPOC in respect to the reference frontends in clean training per SNR

	MFCC	PNCC	SSCH	GBFB	AFE
AVG	26.02	-4.94	53.84	15.67	-8.97
-5 dB	155.57	-3.50	1035.42	133.40	-28.10
0 dB	190.81	-6.91	729.72	159.94	-20.47
5 dB	93.63	-5.89	238.28	56.60	-11.60
10 dB	32.19	-5.65	91.59	10.42	-7.11
15 dB	7.70	-4.87	29.31	-1.37	-5.25
20 dB	-0.50	-4.44	5.12	-3.88	-4.64
clean	-3.09	-3.33	-3.13	-3.55	-3.29

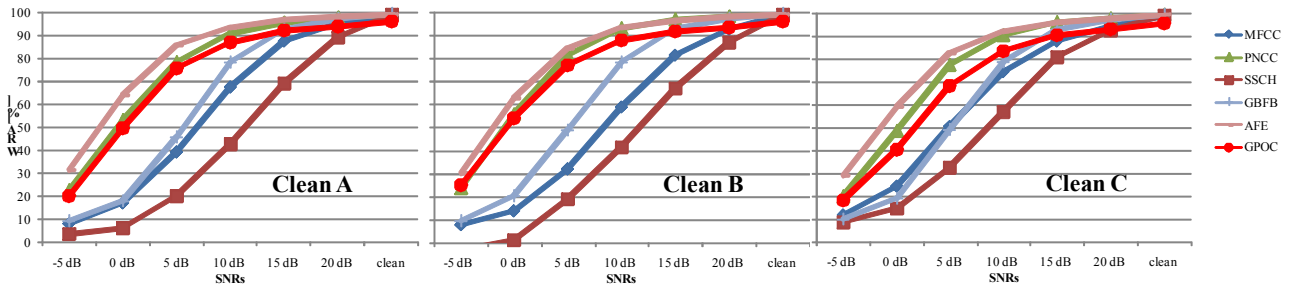

Fig. 6 – Comparison of WRA of the GPOC algorithm and the reference frontends as a function of SNR in clean and multicondition training modes, averaged for the three test sets (A, B and C).

8. DISCUSSION

The results show that the GPOC algorithm has improved robustness to noise when compared to the MFCC, SSCH, and GBFB reference algorithms. This innate noise-robustness of GPOCs is due to both the invariance of the dominant power-flow orientation in the spectrogram to the spectral coloring introduced with noise, as well as the averaging effect of convolving the auditory spectrogram with the 2D kernel bank. While the latter effectively suppresses the present noise, it also damages the high frequency spectro-temporal content of the speech signal, which leads to the algorithm's observed reduction in performance for high SNRs. When compared to the state-of-the-art AFE and PNCC, GPOC only approaches their performance, but it should be emphasized that they explicitly include noise reduction techniques in the feature extraction process.

9. CONCLUSION

The proposed Gaussian Power flow Orientation Coefficients are a novel set of features that describes the pattern of power flow in the auditory spectrogram. The introduced Power flow Orientation Spectrogram is a new representation of the speech signal that captures both the movement of formants, as well as the occurrence of plosives. A big advantage of the GPOC algorithm is that it eliminates the need of feature dimensionality reduction, which is necessary in the conventional spectro-temporal approach. The algorithm shows promising results for various noise types at SNRs below 15 dB, outperforming some of the reference frontends. On the other hand, GPOC's performance falls behind current state-of-the-art, necessitating further improvement before it can be deployed in a real world ASR.

9. REFERENCES

[1] T. Virtanen, R. Singh and B. Raj, eds., *Techniques for Noise-Robustness in Automatic Speech Recognition*. Chichester, UK: John Wiley & Sons, Ltd, Nov. 2012.

[2] B. Gajic and K. K. Paliwal, "Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms", in *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 14 No. 2, Mar 2006

[3] A. Qiu, C. E. Schreiner, and M. A. Escabi, "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," *J. Neurophysiol.*, vol. 90, no. 1, pp. 456–476, 2003.

[4] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proceedings of Interspeech 2002*, 2002, pp. 25–28.

[5] M.R. Schädler, B.T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", *J. Acoust. Soc. Am.* Volume 131, Issue 5, pp. 4134-4151, 2012.

[6] Q. Wu, L. Zhang, G. Shi, "Robust Multifactor Speech Feature Extraction Based on Gabor Analysis", in *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 19 Iss. 4, May 2011.

[7] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," *Apple Technical Report #35*, Apple Computer Library, Cupertino, CA 95014, 1993.

[8] J. Sung, S. Y. Bang, and S. Choi, "A Bayesian network classifier and hierarchical Gabor features for handwritten numeral recognition," in *Pattern Recognition Letters*, vol. 27, no. 1, pp. 66-75, 2006.

[9] H. G. Hirsch and D. Pearce, The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition, *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, France, 2000.*

[10] C. Kim, and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", in *Proc. ICASSP*, pp. 4574–4577, 2010.

[11] ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Frontend feature extraction algorithm; Compression algorithm", *ETSI ES 201 108 v1.1.3 (2003-09)*, Sep 2003.