PQS 2016 5th ISCA/DEGA Workshop on Perceptual Quality of Systems
29-31 Aug 2016, Berlin, Germany

Non-intrusive Estimation of Noisiness as a Perceptual Quality Dimension of Transmitted Speech

Friedemann Köster, Gabriel Mittag, Tim Polzehl, Sebastian Möller

Quality and Usability Lab
Technische Universität Berlin, Germany
friedemann.koester@tu-berlin.de, gabriel.mittag@gmail.com,
tim.polzehl@tu-berlin.de, moeller@tu-berlin.de

Abstract

This article presents a new approach to the non-intrusive quality estimation of transmitted speech. Traditional estimation methods exhibit limitations to providing diagnostic information and for practical monitoring purposes. The new approach merges solutions to overcome the existing limitations and intends to provide a new user-friendly estimator. We present an overview and the planned structure of the proposed model. In order to provide diagnostic information, the method of assessing perceptual quality-relevant dimensions is applied. One of these quality dimensions is *Noisiness*, which describes degradations like background noise, circuit noise, or coding noise. As a fundamental component of the proposed model, a non-intrusive parametric *Noisiness* estimator is presented. The estimator is based on nine different features extracted from the output signal only. Using a linear regression, the features are mapped onto the *Noisiness*. The *Noisiness* estimator is trained on two and tested on three individual subjective databases. In addition, the performance of the resulting estimator is compared to the diagnostic intrusive estimator DIAL (Diagnostic Intrusive Assessment of Listening quality). The results prove that the presented estimator provides high reliability and indicate the applicability and value for non-intrusive diagnostic quality estimation.

Index Terms: speech transmission quality, non-intrusive, diagnostic information, evaluation methods, noisiness

1. Introduction

Within a telephone services, the quality of a transmitted speech signal can be affected, and thus also be degraded. The network and terminal devices responsible for this, referred to as *quality elements* [1], are codecs, packet loss, noise, linear and non-linear distortions and others [2]. In this regard, the assessment of the quality experienced by service users, the QoE (*Quality of Experience*) [3], is of major interest. Traditional methods for the QoE assessment of transmitted speech are subjective experiments with naïve participants resulting in the MOS (*Mean Opinion Score*) [4]. Since subjective methods demand a significant effort to prepare and conduct the studies, instrumental signal-based models have been established. They can be divided into two groups based on the input information they require [5]:

- *Intrusive* methods compare a reference signal with a degraded signal and map the differences to a predicted MOS rating.
- *Non-intrusive* methods map specific signal characteristics of the degraded signal to a predicted MOS rating.

Regarding intrusive models, the long-term standard for NB (*narrowband*: 300 - 3,400 Hz) and WB (*wideband*: 50 - 7,000 Hz) transmission channels recommended by the ITU-T (*International Telecommunication Union*) has been the PESQ (*Perceptual Estimation of Speech Quality*) model for narrowband and WB-PESQ for wideband [6, 7]. It has now been replaced by its successor POLQA (*Perceptual Objective Listening Quality Assessment*), that also considers SWB (*super-wideband*: 50 - 14,000 Hz) transmissions [8]. These models proved to provide reliable estimations of the overall quality. However, intrusive models exhibit certain inherent limitations when put in practice [9]:

- If the resulting MOS estimation is suboptimal it provides little insight into the reason for the quality-loss. More precisely, two different degraded speech stimuli (e.g. one degraded by background noise and the other by clipping) can both be rated with the same (low) MOS value. No diagnostic information is provided.
- The intrusive models demand that the non-degraded input signal of the transmission channel is available, making them only valuable in a laboratory context. For practical online monitoring applications only non-intrusive models that do not demand the input signal are useful.

We present an approach that overcomes both limitations. In order to do this, a non-intrusive speech quality estimator based on perceptual dimensions for diagnosing transmitted speech is presented. After a short overview of related work and an introduction to perceptual dimensions, the new approach is introduced. As a fundamental part of the proposed model a non-intrusive estimator for the perceptual dimension *Noisiness* and its results will be shown. The performance of the presented estimator is discussed and conclusions regarding a non-intrusive overall speech quality estimator are drawn.

2. Related Work

The two aforementioned limitations have been part of various studies. It was investigated how to provide diagnostic information to give insights into quality-losses and furthermore non-intrusive models for estimating the overall quality have been developed.
To obtain diagnostic information the approach of assessing quality-relevant *perceptual dimensions* is applied. The idea underlying this method is the following one: Using a telephony service, the listener will be presented with a sound event. This event might be degraded by the aforementioned quality elements and causes a perceptual event within the listener. The per-

http://dx.doi.org/10.21437/PQS.2016-16

ceptual event is of a multidimensional nature and is composed of explicit perceptual *features*. The features are connected to specific quality elements and can be described by attributes such as loudness or timbre [1]. If the perceptual space is Cartesian and each feature lies along one of the orthogonal axes, the features are referred to as perceptual dimensions [10]. In [11], four perceptual dimensions for NB and WB transmission channels were identified:

- *Noisiness*: describes degradation such as background noise, circuit noise or coding noise. It is labeled with "not noisy" and "noisy".
- *Discontinuity*: describes degradations concerning isolated or non-stationary distortions introduced by e.g. the loss of packets. It is labeled with "continuous" and "discontinuous".
- *Coloration*: describes degradation resulting from frequency response distortions, introduced by e.g. bandwidth limitations. It is labeled with "uncolored" and "colored".
- *Loudness*: is important for the overall quality and the intelligibility. It is labeled with "optimum sound level" and "non-optimum sound level".

The resulting quality profile facilitates the mapping of the overall quality on the basis of perceptual dimensions giving the possibility to identify reasons for quality losses. Also, [12] showed that it is possible to directly quantify the identified perceptual dimensions in a subjective test. The proposed subjective method is similar to what is recommended for noisy speech signals in ITU-T Rec. P.835 [13]. Yielding from these subjective experiments, the intrusive model DIAL (*Diagnostic Instrumental Assessment of Listening quality*) that predicts the overall quality as well as the introduced perceptual dimensions has been developed [5]. This model serves as reference for the proposed *Noisiness* estimator.

Over the past years, non-intrusive models gained more attention for telephony service providers. Since the input speech signal of a transmission channel is mostly not readily available, intrusive models are not useful for the online monitoring purposes. This is, however, the main goal service providers wish to achieve when talking about instrumental models and service evaluation. To provide new models, ITU-T performed a competition to standardize a non-intrusive method in 2004 that produced two submissions. One is the now recommended standard ITU-T P.563 [14]. The algorithm generates an internal reference as replacement for the missing input signal using LPC-analysis and showed to be reliable for NB telecommunication scenarios. The second is called ANIQUE (*Auditory Non-Intrusive QUality Estimation*) and uses the approach of modeling the representation of the speech signal at the central level of the human auditory system. The temporal envelope representation of the speech is used for that [15]. However, both algorithms are only recommended for NB speech transmission and provide no diagnostic information. Currently, ITU-T launched a new standardization process to provide a non-intrusive model that is also suitable for WB and SWB speech transmission [16].

In addition to the non-intrusive overall quality estimators, models to estimate subjective data collected with the ITU-T Rec. P.835 test paradigm have been developed. The 3Quest (*3-fold Quality Evaluation of Speech in Telecommunications*) [17] model predicts the quality of the speech signal, the background noise, and the overall quality. Apart from that, the ITU-T is currently working on the work-item P.ONRA (*Perceptual Objective Noise Reduction*) [18] that is intended to be used as a new standard method to predict ITU-T Rec. P.835 scores.

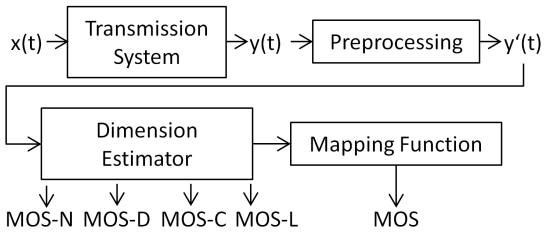

Figure 1: *Structure of the new non-intrusive estimator. N-Noisiness, D-Discontinuity, C-Coloration, L-Loudness.*

3. Proposed Non-intrusive Speech Quality Estimator

To overcome the aforementioned limitations of current estimators a new approach of non-intrusive quality prediction is suggested. To this end, diagnostic information, WB and SWB transmission channels as well as practical monitoring operation will be analyzed. A drafted overview of the planned model structure can be seen in Figure 1. Essentially, the structure consists of three main blocks:

- Preprocessing: here the signals will be preprocessed by filtering and level alignments as well as separation of active and non-active segments via *Voice Activity Detection* (VAD).
- Dimension estimators: this block actually consists of four sub-blocks. Each block will be one estimator for one of the four perceptual dimensions, D-*Discontinuity*, N-*Noisiness*, L-*Loudness*, C-*Coloration*.
- Mapping function: in this block the separate estimations of the perceptual dimensions will be mapped to the overall speech quality using their coherency described in [12].

The aim of this estimator is to provide a new non-intrusive model that provides diagnostic information. For this, dimension estimators must be developed. The concept of these estimators is to extract interpretable parameters of the output speech signal that can be mapped to the corresponding perceptual dimension. As a first step towards putting a diagnostic estimator into practice, an estimator for the dimension *Coloration* has been developed and showed to provide reliable results [19]. Based on these results, as the next step towards putting the estimator into practice, the dimension estimator for *Noisiness* is presented in the next sections.

4. Evaluation Data

For the evaluation of the *Noisiness* estimator, five databases are available. They consist of German speech samples with different sentences (double sentences, duration: 8 s - 12 s) and speakers (4 - 12) as well as subjective ratings for the perceptual dimensions gathered following the paradigm presented in [12]. Thus, the evaluation is based on subjective *Noisiness* ratings on double-sentences and condition base.
The first two databases *DAT1* and *DAT2* were created to identify the perceptual dimensions in [12]. The other three databases *Swiss01*, *Swiss02* and *Swiss03* were used to validate

Table 1: *Overview of the five databases.*

Database	Conditions	Stimuli	Speaker	Hours of Speech
	Training-Data			
$DAT1$	66	792	12	$\approx$ 1.76 h
$Swiss02$	50	200	4	$\approx$ 0.44 h
	Test-Data			
$Swiss01$	50	200	4	$\approx$ 0.44 h
$Swiss03$	54	216	4	$\approx$ 0.48 h
$DAT2$	76	912	12	$\approx$ 2.03 h

the POLQA model [5]. The databases are mixed-band (NB, WB, and SWB) and contain signal-correlated, uncorrelated as well as ambient background noise, temporal clipping, different codecs, packet loss, different frequency distortions, and combinations of these degradations. With these two sets we have data that covers both, simulated and live-recorded degradation. We decided to mix the data and use two databases (*Dat1* and *Swiss02*) for training and three databases (*DAT2*, *Swiss01*, and *Swiss03*) for testing. An overview of the databases can be seen in Table 1.

5. Features to Estimate the Perceptual Dimension Noisiness

Features that can be extracted from the output signal have to be determined to develop a non-intrusive *Noisiness* estimator. To achieve this, the signal is first divided into time segments with fixed length, for each segment the VAD then decides whether it has speech activity or is inactive. To find the best suitable features for a linear regression model we applied a *repeated sequential cross-validation feature selection*. To this end, all training databases (*DAT1* and *Swiss02*) are merged and then divided into one third training and two thirds test-sets (k-cross-validation $K = 3$). Using the training set, the method sequentially selects features ($x \in P$) to best fit a linear regression model:

$$\widehat{MOS_N} = \hat{f}_m(P) = \beta_0 + \sum_{j=1}^{J} \beta_j x_j \tag{1}$$

The methods adds features as long as the *Pearson* correlation ρ [20] between the estimated *Noisiness* MOS values $\widehat{MOS_N}$ and the subjective *Noisiness* MOS values MOS_N is increasing:

$$\rho(MOS_N, \widehat{MOS_N}) = \frac{1}{K} \sum_{k=1}^{K} \rho(MOS_N^{(K)}, \widehat{MOS_N^{(K)}}) \tag{2}$$

Reaching a determined threshold, the algorithm stops. This method is repeated 100 times with varying training and test sets (cross-validation) making sure a wide variety of combinations of the samples are processed. Finally, the algorithm indicates the features most used and the mean correlation for the training sets. Applying this procedure, nine final features were identified and are outlined in the following:

Noise Level (NL): Mean noise level - describes the intensity of the noise level for background noise in non-active segments depending on the noise robust VAD. As can be seen in Equation 3, the NL feature is calculated from the average logarithmic PSD (*Power Spectral Density*) $\hat{\phi}_{nn,n}(\mu)$ of all frequency segments μ, weighted with $H_{ph}(\mu)$ which is the A weighting curve that follows the ANSI S1.42 standard [21].

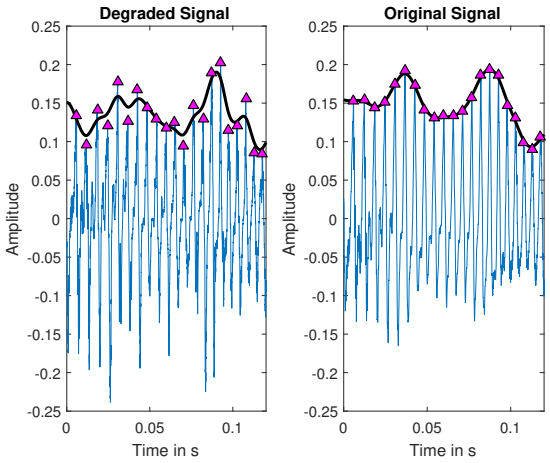

Figure 2: *Signal, Pitch Peaks and Spline Envelope of a Noisy (left) and a Noise Free signal (right).*

$$NL = 10\,log\left(\frac{1}{N}\sum_{\mu=1}^{N}\hat{\phi}_{nn,n}(\mu)H_{ph}(\mu)\right) \tag{3}$$

Pitch Envelope Distortion (PED): This feature assesses signal-correlated disturbances caused by multiplicative noise in active segments. It is assumed, that in clean signals the residual between the peaks and its smoothed envelope is smaller than in noisy signals (compare Figure 2). Consequently, the peaks of the envelope ($y_a(k)$) are found and a cubic smoothing spline ($spy_a(k)$) [22] is fitted to them. The average of the normalized residuals over K peaks then gives the PED for each time segment l:

$$f(l) = \frac{1}{K}\sum_{k=1}^{K}\frac{|spy_a(k) - y_a(k)|}{|y_a(k)|} \tag{4}$$

Frequency Variation (FV): Speech signals with strong noise degradation often have a flat noise floor in the frequency domain. An example of this effect is shown in Figure 3 where the logarithmic PSD of the red noisy signal has a higher variation than the blue unimpaired signal. In order to apply this effect, the sum of the absolute differences of the logarithmic PSD ($P_{yy}(\mu)$) is calculated. Thereby only the values for frequencies within the band-limits are considered. The mean over all inactive segments then gives the frequency variation of the logarithmic PSD:

$$FV = \sum_{\mu=\mu_{1+1}}^{\mu_2} abs\Big(P_{yy}(\mu) - P_{yy}(\mu-1)\Big) \tag{5}$$

Time Variation (TV): This feature uses the sum of the absolute differences of the speech signal in the time domain. The mean of all inactive segments then gives the time variation.

$$TV = \sum_{k=k_{1+1}}^{k} abs\Big(y_i(k) - y_i(k-1)\Big) \tag{6}$$

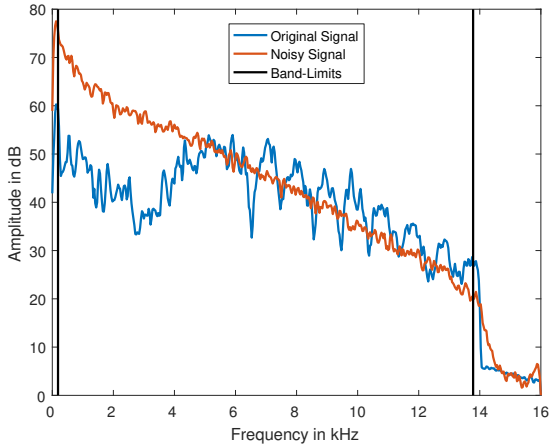

Figure 3: *PSD of a noisy (red) and a noise-free (blue) signal.*

Cepstrum Statistics: The following five features are statistical values extracted from the signals cepstrum [23]. These statistics have shown to be suitable for further analysis of the speech signal [14]. For active segments the standard deviation **CEP-STDa** and the skewness **CEP-SKEWa** are extracted. For inactive segments the standard deviation **CEP-STDi**, the skewness **CEP-SKEWi**, and the kurtosis **CEP-KUTi** are extracted. The statistics are calculated for each 5 ms long segment and the mean of all segments then results in the respective feature.

6. Noisiness Estimator

Applying the feature selection algorithm on the training databases (see Equation 1) enables us to identify the most important features with a positive influence on the correlation, and to eliminate the ones with a negative influence on the correlation. In addition, the feature selection algorithm allows us to extract the optimal coefficients for a linear regression model to best map the perceptual dimension *Noisiness*. Adapted from the results of the selection method, we developed a *Noisiness* estimator based on the 9 introduced features and the corresponding regression coefficients (each making a significant contribution, $p < 0.05$):

$$\begin{aligned}\widehat{MOS_N} = {} & 2.86 - 7.98 \cdot 10^{-4} \cdot \mathbf{NL}^2 \\ & - 41.78 \cdot \mathbf{PED} - 1.28 \cdot \mathbf{FV} \\ & - 1.05 \cdot 10^{3} \cdot \mathbf{TV} - 0.14 \cdot \mathbf{CEP\text{-}STDa} \\ & - 0.19 \cdot \mathbf{CEP\text{-}SKEWa} - 5.50 \cdot \mathbf{CEP\text{-}STDi} \\ & - 0.06 \cdot \mathbf{CEP\text{-}SKEWi}^2 - 0.07 \cdot \mathbf{CEP\text{-}KUTi}\end{aligned} \quad (7)$$

The *Noisiness* estimator is now applied on the test data separately and jointly (*Swiss01*, *Swiss03* and *DAT2*). The results can be seen in Table 2 and in Figure 4. The results show that the estimation of the perceptual dimension *Noisiness* provides stable results. For all three databases correlations of 0.90 and above as well as RMSEs (*Root Mean Square Error*) below 0.48 are reached.

The nine extracted features seem to cover both, background noise and signal-correlated noise (that are both present in the test data), with a high reliability. Especially the *Swiss01* and the *DAT2* databases show high correlations (0.91) with few outliers (compare Figure 4, red crosses for *Swiss01* and pink diamonds for *DAT2*). On the other hand, the estimator seems to slightly

Table 2: *Results of the Noisiness estimator and the DIAL model.*

Database	Proposed Estimator Correlation ρ	Proposed Estimator RMSE	DIAL Correlation ρ	DIAL RMSE
$Swiss501$	0.91	0.45	0.82	1.29
$Swiss503$	0.90	0.48	0.86	0.39
$DAT\ 2$	0.91	0.37	0.68	0.56
all	0.90	0.43	0.78	0.75

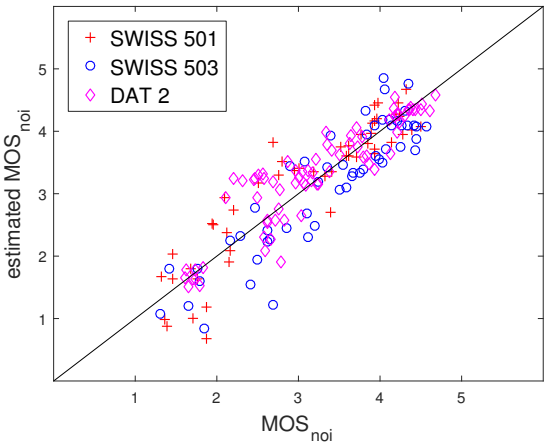

Figure 4: *Results of the Noisiness estimator.*

under-predict the subjective *Noisiness* ratings for the Swiss03 data between MOS values of 1 and 3 (blue dots).

The proposed model outperforms the DIAL model for all three test databases in terms of correlation and in terms of RMSE for two out of three databases (compare Table 2). For database Swiss03 the proposed model achieves a higher correlation, but also a higher RMSE than the DIAL model. Again, this can be seen in Figure 4 where low Coloration ratings are slightly under-estimated (blue crosses). Overall, considering the different data and the non-intrusive approach, the results show that the proposed estimator is capable of predicting the perceptual dimension Noisiness.

7. Conclusion and Outlook

We presented a new approach towards the non-intrusive estimation of speech quality. While a first estimator for the perceptual dimension *Coloration* has already been developed, as a second step a *Noisiness* estimator and its corresponding features are introduced. The evaluation of the estimator shows that it provides correlations and errors to a high degree of reliability. Respecting the complications of a non-intrusive approach on the one hand, as well as the number of test databases included on the other hand, the proposed estimator builds a fundamental part towards a integrated non-intrusive quality estimator.
In future work we will introduce non-linear regression approaches that may reduce the error and increase the correlation. As a further step, estimators for discontinuity and loudness are needed. Having these predictors at hand we strive to map the overall quality, resulting in a non-intrusive speech quality estimator.

8. Acknowledgements

The presented work was supported by the Federal Ministry of Education and Research, Germany (01IS12056) and the Software Campus.

9. References

[1] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Berlin: Springer Science & Business Media, 2005.

[2] A. Raake, *Speech Quality of VoIP Assessment and Prediction*. Chichister, West Sussex: John Wiley & Sons, 2006.

[3] S. Möller and A. Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*. Berlin: Springer, 2014.

[4] ITU-T Recommendation. P.800, "Methods for subjective determination of transmission quality," 1996.

[5] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Berlin: Springer, 2011.

[6] ITU-T Recommendation. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," 2001.

[7] ITU-T Recommendation. P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," 2007.

[8] ITU-T Recommendation. P.863, "Perceptual Objective Listening Quality Assessment," 2011.

[9] F. Köster, S. Möller, J.-N. Antons, S. Arndt, D. Guse, and B. Weiss, "Methods for Assessing the Quality of Transmitted Speech and of Speech Communication Services," *Acoustics Australia*, vol. 42, no. 3, pp. 179 – 184, December 2014.

[10] I. Borg and P. Groenen, *Modern Multidimensional Scaling - Theory and Applications*, 2nd, Ed. New York, NY: Springer Series in Statistics, 2005.

[11] M. Wältermann, A. Raake, and S. Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission." Acta Acustica united with Acustica, 2010, pp. 1090–1103.

[12] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*. Berlin: Springer, 2012.

[13] ITU-T Recommendation. P.835, "Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Supression Algorithm," 2003.

[14] ITU-T Recommendation. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," 2004.

[15] D. S. Kim and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment." Montreal, Canada: Proc. IEEE ICASSP, 2004.

[16] ITU-T Temporary Document TD (GEN/14), *Technical Requirement Specification P.SPELQ*, 2014.

[17] HEAD acoustics GmbH, "3QUEST:3-fold Quality Evaluation of Speech in Telecommunications," 2008.

[18] ITU-T Contribution COM 12-101, "P.ONRA Requirement Specification, Draft V03," 2013.

[19] G. Mittag, F. Köster, and S. Möller, "Non-intrusive Estimation of the Perceptual Dimension Coloration," in *Fortschritte der Akustik, DAGA 2016: Plenarvortr. u. Fachbeitr. d. 42. Dtsch. Jahrestg. f. Akust.* DEGA, 2016.

[20] J. Bortz, *Statistik: fr Human- und Sozialwissenschaftler (Springer-Lehrbuch) (German Edition)*, 6th ed. Springer, 9 2004. [Online]. Available: http://amazon.com/o/ASIN/354021271X/

[21] A. S1.42, *Design Response of Weighting Networks for Acoustical Measurements*, American National Standard, 2001.

[22] C. de Boor, *A Practical Guide to Splines*. New York: Springer, 1978.

[23] K. Steiglitz and B. Dickinson, "Computation of the complex cepstrum by factorization of the z-transform," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '77.*, vol. 2, May 1977, pp. 723–726.