A ROBUST FRONTEND FOR ASR: COMBINING DENOISING, NOISE MASKING AND FEATURE NORMALIZATION

Maarten Van Segbroeck and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab,
University of Southern California, Los Angeles, USA
{maarten, shri}@sipi.usc.edu

ABSTRACT

The sensitivity of Automatic Speech Recognition (ASR) systems to the presence of background noises in the speaking environment, still remains a challenging task. Extracting noise robust features to compensate for speech degradations due to the noise, regained popularity in recent years. This paper contributes to this trend by proposing a cost-efficient denoising method that can serve as a preprocessing stage in any feature extraction scheme to boost its ASR performance. Recognition performance on Aurora2 shows that a noise robust frontend is obtained when combined with noise masking and feature normalization. Without the requirement of high computational costs, the method achieves similar recognition results when compared to other state-of-the art noise compensation methods.

***Index Terms*—** speech enhancement, noise robust feature extraction, speech recognition

1. INTRODUCTION AND PRIOR WORK

Over the years, much effort has been devoted on developing techniques for noise robust Automatic Speech Recognition (ASR). Besides the variability in their approaches, all these techniques have as common goal to make the ASR system more resistant to the mismatch between training and testing conditions. Noise reduction techniques can be applied at different levels of the ASR-system: (i) speech enhancement at the signal level [1, 2, 3], (ii) robust feature extraction [4, 5, 6, 7] or (iii) adapting the back-end models [8, 9, 10].

In real-life situations, the statistics of the background noise are not known beforehand and difficult to predict. Hence, most appealing are those techniques that do not rely on important assumptions about the noisy conditions or on parameters that need to be trained intensively to perform well under specific noise (and speech) scenarios. The aim of extracting noise robust features should be to make only weak or no assumptions about the noise. This is a strong argument in favour for the ongoing and recent research for finding a representation that is insensitive to a wide range of noise distortions when applied to an ASR, e.g. bottle-neck features [11], Power-Normalized coefficients [12], Gabor features [13], Gammatone Frequency Cepstral Coefficients [14], to name only a few.

The work in this paper was motivated by the study presented in [15]. It was shown that a computationally efficient frontend implementation could achieve similar recognition performance as computationally expensive techniques such as Parallel Model Combination (PMC) [16] and Missing Data Techniques (MDT) using data imputation during decoding [17].

This paper contributes to the ongoing research on noise robust ASR by proposing a combined application of robust feature extraction, feature normalization and model adaptation on speech that has been *denoised* by a speech enhancement technique taking into account the voicing characteristics of the speech.

An important cue to detect, measure and extract speech information - even in extreme noisy conditions - is the presence of a fundamental frequency in the human voice (pitch) and its corresponding spectral harmonics. This fundamental speech property is exploited in the proposed noise suppression algorithm. Here, the spectrum of the background noise is estimated from the residual signal obtained after removal of the harmonic spectral peaks arising from the voice, which is then used to suppress the noise in noisy speech. Unlike other speech enhancement approaches, such as Wiener Filtering [2] or Spectral Subtraction [1], the proposed method does not require a speech activity detector or assumes stationarity of the noise over a relatively large time window and is able to reduce unwanted speech degradations by the limited leakage of voicing energy in the spectral subbands of the noise prior to subtraction.

The performance of the presented speech enhancement method is tested on the Aurora2 benchmark database and recognition results are produced by a HTK-based full digit recognizer [18] and by a speech recognition system built with the IBM Attila toolkit [19] in which context-dependent phone models are used. Accuracy results are presented on a set of different feature representations which are pre-processed by the proposed denoising algorithm in combination with noise masking, and further normalized to compensate for channel and speaker variations.

The outline of the paper is as follows. Section 2 presents the speech enhancement technique taking the voicing characteristics of the speech into account. This technique will serve as pre-processing step for the feature extraction module of section 3, where additional steps are applied to obtain a robust front-end for ASR. The experimental setup and results are described in section 4. Final conclusions and future work are given in section 5.

2. SPEECH DENOISING

2.1. Removal of voicing

During voiced speech periods, the noisy speech is characterized by the presence of strong periodicity arising from pitch and pitch multiples (harmonics). Therefore, the first step to estimate the noise is to remove this periodicity from the noisy speech signal.

To obtain this *unvoiced* noisy signal, the periodicity of the signal arising from the voiced speech is estimated using the harmonic decomposition method proposed in [20]. Here, an initial pitch estimate is computed by a subharmonic summation method where the target pitch value is confined to the frequency range from 50 Hz to 800 Hz. A pitch synchronous framing is subsequently applied on the signal to obtain overlapping *segments* with a length of two pitch periods

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

and a single period of frame-shift. Denoting the noisy speech in the time-domain by $y(t)$, the pitch epoch index by p and the estimate of the double pitch period by Ω_p, the unvoiced noisy signal can then be written as the following subtraction in the time-domain:

$$x_p(n) = y_p(n) - v_p(n) \;\text{ with } 0 \leq n < \Omega_p \qquad (1)$$

where the voiced or harmonic component $v_p(n)$ of the input signal is defined as

$$v_p(n) = \Big(1 + \frac{e_p n}{\Omega_p}\Big) \sum_{k=0}^{K_p} a_{k,p} \cos(\omega_p n) + b_{kp} \sin(\omega_p n) \qquad (2)$$

with $\omega_p = 2\pi f_{0p} n$, $b_{0p} = 0$, f_{0p} the pitch estimate of each segment p, and K_p the number of harmonics in the frequency range from 0 to the Nyquist frequency. The change in amplitude over the Ω_p samples is taken into account by the linear modulation factor $(1 + e_p n/\Omega_p)$.

For each segment p, we choose to estimate $x_p(n)$ by minimizing the Penrose regression function, i.e.

$$\hat{x}_p(n) = \min_{\gamma_p} ||y_p(n) - v_p(n)||^2 \;\text{ with } \gamma_p = \begin{bmatrix} a_{kp} \\ b_{kp} \\ f_{0p} \\ e_p \end{bmatrix} \qquad (3)$$

using the optimization approach described in [20]. After concatenating over time all residual signals of eq. (3), we obtain the unvoiced noisy signal $x(t)$.

2.2. Noise estimation

If the short-term Fourier spectrum of $x(t)$ is given by $X(f, k)$, computed every 10 ms using a 25 ms frame of Hamming windowed data, then the short-time sub-band energy of the noise can be estimated from the minimum statistics of $X(f, k)$. The minimum statistics method prevents the subtracting of high energy unvoiced speech regions present in $x(t)$ without the need of a voice activity detector. In our approach, the absolute value of the noise spectrum is then estimated as

$$|\hat{N}(f, k)| = \tilde{\mathbf{X}}(\alpha(f), k) \qquad (4)$$

where $\tilde{\mathbf{X}}(f, k)$ is a vector containing sorted sub-band energy values over $2\lambda + 1$ frames centralized around each frame k, i.e.:

$$\tilde{\mathbf{X}}(f, k) = \text{sort}\{|X(f, k - \lambda)| \cdots |X(f, k)| \cdots |X(f, k + \lambda)|\}$$

and where $1 \leq \alpha(f)(2\lambda + 1) \leq 2\lambda + 1$ with $\alpha(f)$ a frequency dependent index value that is proportional to the observed noise energy level in frequency subband f.

If we define the log-energy of $x(t)$ and the voiced signal $v(t)$ for each frame by $E_x(k)$ and $E_v(k)$ respectively, then the ratio $V(k) = E_v(k)/E_x(k)$ is a measurement for the voicing contained in the signal. The voicing information can then be integrated in equation (4) to update the noise energy values as follows:

$$|\bar{N}(f, k)| = \rho |\hat{N}(f, k)| \;\text{ with } \rho = \begin{cases} \rho_s, & \text{if } V(k) \geq 1 \\ \rho_n, & \text{if } V(k) < 1 \end{cases}$$

By choosing the parameter ρ_s in the range $[0.5, 1]$ and ρ_n within $[1, 1.5]$, a proper trade-off can be found between noise suppression during noise/unvoiced speech periods and speech degradation during speech. By smoothing the values of $V(k)$ over time, suppression of unvoiced speech frames adjacent to speech frames can be reduced.

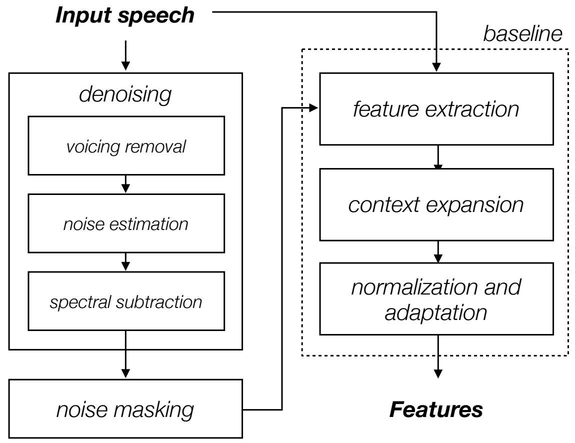

Fig. 1. Overview of pre-processing steps in the proposed frontend to extract robust features. The dashed rectangle denotes the baseline frontend.

2.3. Spectral noise subtraction

In order to obtain a *denoised* version of the noisy signal, we adopt the subtraction rule that was proposed in [21]. The spectral magnitudes of the noise estimate $|\bar{N}(f, k)|$ are subtracted from the spectrum of the noisy signal $|Y(f, k)|$, taking into account an oversubtraction factor that is computed as a function of the signal-to-noise ratio per frequency subband. A spectral floor constant is also defined to set a maximum value for the subtraction. See [21] for more details.

Finally, the denoised speech signal $y_{dn}(t)$ is reconstructed in the time-domain after applying an inverse Fourier Transform on each frame taking into account the (unaltered) phase of the noisy signal $y(t)$ and a division by the values of the Hamming window. Note that a conversion to the time domain is not strictly required when feature extraction would be applied after the denoising stage.

The resulting algorithm is a computational efficient speech enhancement method that can either be applied to improve the signal quality or to boost the performance of ASR systems for a wide range of stationary and non-stationary noise types, even at very low SNR levels.

3. ROBUST NORMALIZED FEATURE SCHEME

3.1. Feature extraction

The proposed denoising algorithm will be applied as a pre-processing step for following feature extraction modules: Mel-Frequency Cepstral Coefficients (MFCC) [22], Perceptual Linear Prediction (PLP) coefficients [23], Gammatone Frequency Cepstral Coefficients (GFCC) [14] and Gabor Features (GBF) [13].

Although other feature representations could have been explored as well, our choice for the above mentioned features was mainly motivated by the popularity of MFCCs and PLPs in ASR systems, the different perceptual characteristics of GFCCs (which are derived from a cochleagram using a Gammatone filterbank), and the psychoacoustically motivated GBF representation (which attempts to model the spectro-temporal processing of the primary auditory cortex by a set of Gabor filters with varying temporal and spectral modulation frequencies [24]).

3.2. Context expansion

In most ASR frontends, context information is typically included by taking the feature values of neighboring frames into account. This can be done either by augmenting the feature streams with their first and second order derivatives or by applying a linear projection, such as linear discriminant analysis (LDA), on feature streams that are constructed by stacking successive frames. Both techniques will be investigated by respectively the HTK and Attila speech recognition system.

3.3. Noise masking and normalization

Robustness is further improved by applying mean and variance normalization (MVN) to compensate for mismatches in linear filtering and dynamic range reduction introduced by both convolutional and additive noise sources. In the case of clean training and multi-style testing, mean normalization will introduce a mismatch in the bias caused by the silence frames in the long-term average. A simple way to compensate for this bias is by applying noise masking, i.e. adding a (white) noise signal with an amplitude relative to the speech level, to both training and denoised test data such that the mismatch in their sub-band energy levels is reduced. In this paper, noise masking was simply done by time domain adding of white noise. Experiments not reported here, have shown that the optimal denoising parameter setting has to be found in combination with this noise amplitude level.

As will be shown in section 4, the combination of denoising with noise masking in the proposed frontend scheme, will not degrade the performance of speech recognized at high SNR levels. In the Attila system, the features are further linearly transformed to normalize out speaker variability by feature-space MLLR (fMLLR) [25].

3.4. Baseline vs. proposed frontend

In our baseline ASR system, the acoustic models are trained and tested on the above mentioned features normalized using MVN and/or fMLLR. This baseline was tested against the proposed frontend which only differs in the use of the proposed denoising algorithm and noise masking, which was also applied on the clean training data. An overview of the proposed ASR frontend is shown in Figure 1.

4. EXPERIMENTAL RESULTS

The evaluation is done on the Aurora2 TI-Digits speech database by two different speech recognition systems that are trained using either the HTK software package or the Attila toolkit. In both cases, the acoustic models are trained on the clean speech training database and tested on the three different noisy test sets of Aurora2.

In the HTK-based system, the digits are modeled as whole word left-to-right HMMs with 16 states per digit and 20 Gaussians with diagonal covariance per state. The acoustic model in the back-end consists of an HMM Gaussian mixture architecture with 16 states per digit and 20 Gaussians per state. The optional inter-word silence is modeled by 1 or 3 states with 36 Gaussians per state, while leading and trailing silence have 3 states. The total number of Gaussians is 3628. The frontend uses either (i) MFCC features, where 23-channel MEL filter bank spectra are transformed to 13-dimensional cepstra, (ii) PLP features with 13 dimensions, (iii) GFCC features, where 64-channel Gammatone filter bank spectra are transformed to 24-dimensional cepstra and (iv) Gabor Features, after applying the critical frequency sampling of [26] and retaining only the temporal modulation frequencies at 0Hz and 2.4Hz as motivated in [27]. MFCC and GFCC features were augmented by dynamic coefficients computed using a window length of 9 frames, to yield respectively 39 and 72-dimensional feature vectors for recognition. Note that for GBF, temporal variations are already integrated by their definition. All features were subsequently mean and variance normalized.

HTK - Aurora2, 8kHz, clean condition training.
Full left-to-right digit HMMs.

Baseline frontend							
Feat.	Test	0dB	5dB	10dB	15dB	20dB	avg.
MFCC	testa	50.4	77.8	91.4	96.6	98.4	82.9
	testb	51.9	79.6	92.9	97.3	98.8	84.1
	testc	50.3	78.4	91.4	96.7	98.4	83.0
PLP	testa	51.7	78.1	91.2	96.3	98.4	83.1
	testb	51.3	79.9	92.7	97.2	98.8	83.9
	testc	52.5	78.9	91.2	96.1	98.3	83.4
GFCC	testa	52.1	80.8	93.2	96.7	98.0	84.1
	testb	59.1	84.3	94.6	97.4	98.4	86.7
	testc	50.1	76.8	90.9	96.0	97.6	82.2
GBF	testa	57.6	80.4	92.0	96.9	98.5	85.0
	testb	60.5	83.0	93.6	97.6	98.8	86.7
	testc	61.2	82.2	92.8	97.0	98.5	86.3
Proposed frontend							
Feat.	Test	0dB	5dB	10dB	15dB	20dB	avg.
MFCC	testa	69.5	88.4	96.1	98.1	98.9	90.2
	testb	69.2	88.2	96.1	98.3	98.9	90.1
	testc	67.0	87.1	94.9	97.8	98.6	89.0
PLP	testa	65.5	86.2	95.2	97.8	98.9	88.7
	testb	63.6	85.3	94.8	98.0	98.9	88.1
	testc	62.4	84.1	94.1	97.5	98.5	87.3
GFCC	testa	67.3	87.1	95.0	97.6	98.3	89.0
	testb	70.4	88.3	95.7	97.8	98.5	90.1
	testc	60.9	82.0	92.2	96.5	97.7	85.8
GBF	testa	69.2	86.8	94.5	97.6	98.6	89.3
	testb	69.3	87.4	95.0	97.9	98.6	89.6
	testc	70.0	86.5	94.5	97.2	98.4	89.3

Table 1. Word recognition accuracy (in %) on the Aurora2 test sets obtained by the HTK system using the baseline and proposed frontend with different types of feature representations.

Table 1 presents the results obtained by the baseline models compared to the full frontend with the speech enhancement algorithm of section 2. Here, all algorithmic tunable parameters where fixed among all feature extraction modules and noise types. Although not extensively tested, a good parameter setting was exper-

System	testa	testb	testc	avg.
SS	80.3	81.8	80.2	80.8
ETSI	87.7	87.1	85.4	86.7
MBFE	89.5	87.5	87.4	88.3
MDT	87.9	89.8	86.6	88.6
proposed	**90.2**	**90.1**	**89.0**	**89.8**

Table 2. Word recognition accuracy (in %) averaged over 0-20dB SNR levels on the Aurora2 test sets achieved with the HTK system using different noise compensation methods.

imentally found by setting $\lambda = 10$, $\alpha(f) = 0.2$, $\rho_s = 0.75$, $\rho_n = 1.25$. For all deployed features, a consistent accuracy improvement was shown at all SNR levels and this is mostly prominent a low SNR conditions 0-5db. Due to the simplicity of the recognition task and the similarity in their results, no general conclusions can be made in the relative performance between the used feature types. Important to notice is that noise masking does not result in performance degradation at high SNR conditions.

The average performance of our method using MFCCs was compared in Table 2 against the spectral subtraction (SS) method of [21], the ETSI advanced frontend (AFE) [28], the Model-Based Feature Enhancement (MBFE) technique of [16] and the Missing Data Theory (MDT) based approach of [17]. In the AFE, noise reduction is done by applying Wiener Filtering, VAD and blind equalization. MBFE exploits Vector Taylor Series approximation to estimate the clean speech from the noisy data from a combined model trained on clean speech and noise. The MDT method estimates reliability masks from noisy data and uses a data imputation technique to reconstruct the missing part of the feature vector. The table shows that the proposed frontend outperforms the ETSI advanced frontend on the 3 test sets and achieves a similar recognition accuracy as the other methods, but with significant less computational complexity.

In the Attila system, context-dependent (CD) models are used to model 19 phones together with 3 phones denoting the silence and the beginning and ending of speech. As in [29], each phone is trained by a 3-state Hidden Markov Model. The Attila Training Recipe (ATR) [19] was followed to train the acoustic models by first initializing the CD models by context-independent models. The CD models are trained on 40 dimensional feature vectors that are derived by applying a LDA transform on a stacked representation of mean variance normalized 13-dimensional PLP features obtained from 9 successive frames. Just like taking first and second order derivative, this approach takes context frames into account but now encoding and decorrelation is applied on the stacked feature vector, which typically results in a slight performance improvement [30].

Finally, feature space MLLR is applied on the data to compensate for speaker variability. Experimental results are given in Table 3. As comparison, the improvement in word accuracy by fMLLR is shown individually to assess the relative sensitivity of the ASR to the degradation caused by the spectral subtraction and noise masking of the proposed frontend. Moreover, unlike the baseline frontend, fMLLR does not degrade the accuracy at 0dB SNR.

5. CONCLUSIONS

A speech denoising algorithm was presented in which noise suppression is achieved by estimating the noise from the residual part of the input signal obtained after removing the periodicity caused by voiced speech. It was shown that when combined with noise masking and feature normalization, the denoising method is an efficient pre-processing step in a robust frontend scheme for the recognition task on the small vocabulary Aurora2 database. When compared to other state-of-the-art methods, similar recognition results are obtained, but at a significant lower computational cost. Future work includes assessing the performance of the denoising on real-life large vocabulary databases and extending the method such that the algorithmic parameters automatically adapt to the observed noise level in each frequency subband.

ATTILA - Aurora2, 8kHz, clean condition training. Context dependent phone models.							
Frontend	Test	0dB	5dB	10dB	15dB	20dB	avg.
Baseline	testa	41.1	66.7	84.5	93.9	97.7	76.7
	testb	41.9	67.1	84.8	94.0	97.7	77.1
	testc	40.4	65.3	82.7	93.1	97.6	75.8
Baseline + fMLLR	testa	39.3	72.1	89.4	96.2	98.1	79.0
	testb	38.3	71.3	90.2	96.1	98.3	78.8
	testc	40.0	69.4	87.8	95.8	98.6	78.3
Proposed	testa	62.9	83.3	93.7	97.3	98.5	87.1
	testb	57.5	81.1	92.2	96.8	98.4	85.2
	testc	56.0	79.3	91.3	96.7	98.1	84.2
Proposed + fMLLR	testa	62.9	83.7	93.9	97.6	98.2	**87.2**
	testb	59.5	83.1	93.2	96.9	98.2	**86.1**
	testc	56.9	81.7	93.3	97.4	98.5	**85.5**

Table 3. Word recognition accuracy (in %) on the Aurora2 test sets obtained by the Attila system using the baseline and proposed frontend with fMLLR.

6. ACKNOWLEDGEMENT

This research was supported by the Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF)

7. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] S. V. Vaseghi and B. P. Milner, "Noise-adaptive hidden markov models based on wiener filters," 1993, pp. 1023–1026.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] H. Hermansky, N. Morgan, A. Bayya, and Ph. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," Genua, Italy, Sept. 1991, pp. 1367–1370.

[5] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Communication*, vol. 12, no. 3, pp. 277–288, July 1993.

[6] B. Kingsbury and S. Greenberg, "The modulation spectrogram: in pursuit of an invariant representation of speech," 1997, pp. 1647–1650.

[7] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, Sept. 2009.

[8] A.P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise," Albuquerque, NM, U.S.A., Apr. 1990, pp. 845–848.

[9] M.F.J. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, UK, Sept. 1995.

[10] J. Droppo, A. Acero, and L. Deng, “Uncertainty decoding with splice for noise robust speech recognition,” in *Proc. ICASSP*, Orlando, Florida, U.S.A., May 2002, pp. 57–60.

[11] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, “Probabilistic and bottle-neck features for lvcsr of meetings,” in *Proc. ICASSP*, 2007.

[12] C. Kim and R. M. R. M. Stern, “Power-normalized coefficients (pncc) for robust speech recognition,” in *Proc. ICASSP*, 2012.

[13] Kleinschmidt M., “Spectro-temporal gabor features as a front end for ASR,” in *Proc. Forum Acusticum Sevilla*, 2002.

[14] Y. Shao, S. Srinivasan, and D.L. Wang, “Incorporating auditory feature uncertainties in robust speaker identification,” in *Proc. ICASSP*, 2002, pp. 277–280.

[15] K. Demuynck, X. Zhang, and H. Van Compernolle, Van hamme, “Feature versus model based noise robustness,” in *Proc. Interspeech*, 2010.

[16] V. Stouten, *Robust automatic speech recognition in time-varying environments*, Ph.D. thesis, K.U.Leuven, ESAT, Sept. 2006.

[17] M. Van Segbroeck, *Robust Large Vocabulary Continuous Speech Recognition using Missing Data Techniques*, Ph.D. thesis, K.U.Leuven, ESAT, Jan. 2010.

[18] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book – version2.2*, Entropic, 1999.

[19] H. Soltau, G. Saon, and B. Kingsbury, “The ibm attila speech recognition toolkit,” in *Spoken Language Technology Workshop (SLT)*, 2010.

[20] H. Van hamme, “Robust speech recognition using cepstral domain missing data techniques and noisy masks,” in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 213–216.

[21] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” in *IEEE Transactions on Speech and Audio Processing*, July 2001, vol. 9, pp. 504–512.

[22] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognitions in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[23] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[24] B. Meyer, S. Ravuri, M.R. Schädler, and N. Morgan, “Comparing different flavors of spectro-temporal features for ASR,” in *Proc. Interspeech*, 2011, pp. 1269–1272.

[25] M. J. F. Gales, “Maximum-likelihood linear transforms for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[26] B. T. Meyer, S. V. Ravuri, M. R. Schadler, , and N. Morgan, “Comparing different flavors of spectro-temporal features for asr,” in *Proc. Interspeech*, 2011, pp. 1269–1272.

[27] T.J. Tsai and N. Morgan, “Longer features: They do a speech detector good,” in *Proc. Interspeech*, 2012.

[28] H.G. Hirsch and D. Pearce, “Applying the advanced etsi front-end to the aurora-2 task,” Tech. Rep. version 1.1, Cambridge University Engineering Department, 2006.

[29] G. Saon, J. M. Huerta, and E.E. Jan, “Robust digit recognition in noisy environments: The ibm aurora-2 system,” in *Proc. Interspeech*, 2001, pp. 629–632.

[30] B. Milner, “Inclusion of temporal information into features for speech recognition,” in *Proc. ICSLP*, 1996, pp. 256–259.