A Subband-Based Stationary-Component Suppression Method Using Harmonics and Power Ratio for Reverberant Speech Recognition

Byung Joon Cho, Haeyong Kwon, Ji-Won Cho, *Student Member, IEEE*, Chanwoo Kim, *Member, IEEE*, Richard M. Stern, *Fellow, IEEE*, and Hyung-Min Park, *Senior Member, IEEE*

***Abstract*—This letter describes a preprocessing method called subband-based stationary-component suppression method using harmonics and power ratio (SHARP) processing for reverberant speech recognition. SHARP processing extends a previous algorithm called Suppression of Slowly varying components and the Falling edge (SSF), which suppresses the steady-state portions of subband spectral envelopes. The SSF algorithm tends to over-subtract these envelopes in highly reverberant environments when there are high levels of power in previous analysis frames. The proposed SHARP method prevents excessive suppression both by boosting the floor value using the harmonics in voiced speech segments and by inhibiting the subtraction for unvoiced speech by detecting frames in which power is concentrated in high-frequency channels. These modifications enable the SHARP algorithm to improve recognition accuracy by further reducing the mismatch between power contours of clean and reverberated speech. Experimental results indicate that the SHARP method provides better recognition accuracy in highly reverberant environments compared to the SSF algorithm. It is also shown that the performance of the SHARP method can be further improved by combining it with feature-space maximum likelihood linear regression (fMLLR).**

***Index Terms*—Harmonics, precedence effect, reverberation, robust speech recognition.**

I. Introduction

NOISE robustness remains an important issue in the field of automatic speech recognition (ASR), because the performance of most ASR systems is seriously degraded when there are differences between training and testing environments. Although many algorithms have been proposed to compensate for these mismatches (e.g., [1], [2]), they are mainly focused on coping with additive noise. Speech in rooms is frequently corrupted by reverberation because it incurs multiple reflections from the rooms' surfaces. Because direct dereverberation in the time domain can be computationally costly (e.g., [3], [4]), subband-based approaches are considered.

Manuscript received August 05, 2015; accepted March 31, 2016. Date of publication April 15, 2016; date of current version April 25, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Frederic Bechet.

B. J. Cho, H. Kwon, J.-W. Cho, and H.-M. Park are with the Department of Electronic Engineering, Sogang University, Seoul 04107, South Korea (e-mail: hpark@sogang.ac.kr).

C. Kim is with the Google Corporation, Mountain View, CA 94043 USA.

R. M. Stern is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2554888

While the human auditory system is more sensitive to modulation frequencies less than 20 Hz (e.g., [5]), very slowly changing components (e.g., less than 5 Hz) are usually produced by noise sources (e.g., [6]). Thus, researchers have tried to improve ASR performance by performing high-pass or band-pass filtering of subband power on a frame-by-frame basis (e.g., [7]). Recently, Kim and Stern proposed a processing method called Suppression of Slowly varying components and the Falling edge (SSF) that accomplishes onset enhancement and steady-state suppression by applying a type of high-pass filtering to the frame-by-frame power of signals that had been passed through a bank of gammatone filters [6]. They demonstrated that SSF processing can achieve significant improvements in ASR performance in reverberant environments.

Although SSF processing can improve robustness of ASR systems, the power contours of processed signals for clean and reverberated speech are still different. The major difference occurs in processing reverberated voiced speech with high power contours, because the power contours of reverberated voiced speech are more smeared over time than those of clean voiced speech. In addition, in reverberant environments, SSF processing may inappropriately remove useful features for recognizing unvoiced phonetic segments with energy concentrations at high frequencies such as fricatives because a high level of energy in a particular channel in previous frames may cause over-subtraction in the current frame.

To overcome these undesirable properties of SSF processing, we present a preprocessing method based on stationary-component suppression in the subband domain, which we refer to as subband-based stationary-component suppression method using harmonics and power ratio or "SHARP" processing. The useful features of unvoiced speech are retained by detecting the frames that contain them based on energy distributions across frequency, and the degree of subtraction is reduced for these frames. To more closely match the power contours of clean and reverberated voiced speech, the floor value in the subtraction is boosted to stretch the power contours along the time axis, using a measure of harmonicity to detect voiced-speech frames. In addition, the combination of SHARP processing with feature-space maximum likelihood linear regression (fMLLR) [8] that is known to be effective in achieving speaker adaptation is demonstrated to provide improved robustness in reverberant environments.

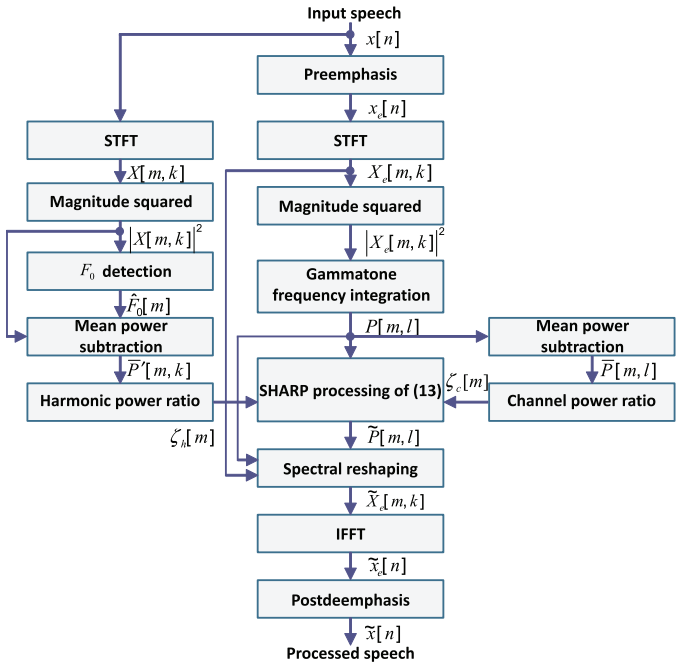

Fig. 1. Summary of the SHARP processing procedure.

II. SHARP Processing

Fig. 1 shows the overall SHARP processing procedure. A short-time Fourier transform (STFT) is performed using a 50-ms Hamming window with a 10-ms frame shift.[1] Magnitude-squared STFT outputs are used to obtain the power $P[m,l]$ at the mth frame and lth gammatone channel as in [6] and [9]:

$$P[m,l] = \sum_{k=0}^{N/2} |X_e[m,k]H_l[k]|^2, \quad 0 \le l \le L-1 \qquad (1)$$

where N and L denote the discrete-Fourier-transform size and the number of gammatone channels, respectively. $X_e[m,k]$ is the signal spectrum at the mth frame and kth frequency bin, and $H_l[k]$ is the transfer function of the lth channel evaluated at the kth frequency bin, in a gammatone filter bank whose center frequencies are linearly spaced in equivalent rectangular bandwidth (ERB) [10] between 200 Hz and 8 kHz.

The power $P[m,l]$ is low-pass-filtered to obtain $M[m,l]$ by

$$M[m,l] = \lambda M[m-1,l] + (1-\lambda)P[m,l] \qquad (2)$$

where λ denotes a forgetting factor. In SSF processing [6], [9], the processed power is obtained by

$$\tilde{P}[m,l] = \max(P[m,l] - M[m,l], c_0 M[m,l]) \qquad (3)$$

where c_0 is a small fixed coefficient to set the floor value. Because $M[m,l]$ is subtracted from $P[m,l]$, $\tilde{P}[m,l]$ is essentially a high-pass-filtered signal with suppression of slowly varying components and a falling edge in its power contour.

As an illustrative example, Fig. 2 shows the power spectra of clean and reverberated speech using SSF processing of (3). The original power spectra without any processing are also shown.

[1]A longer-duration window is used because "medium-time" processing is more effective for noise estimation or compensation [6], [9].

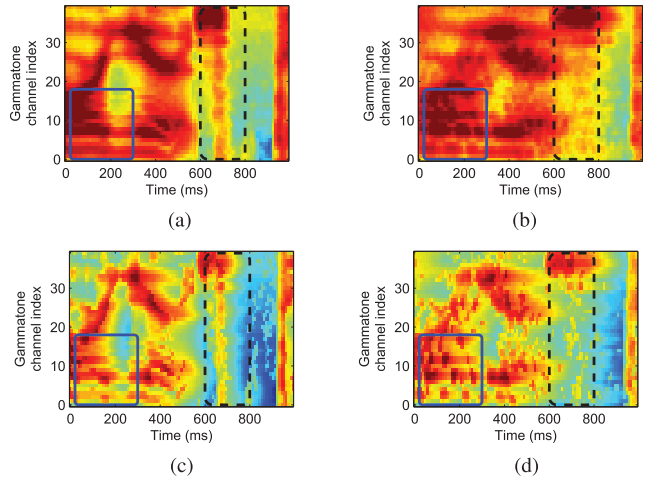

Fig. 2. Power spectra of clean and reverberated speech in gammatone channels processed using either no or SSF processing. The reverberation time RT_{60} used to generate the reverberated speech was 1.2 s. The values are depicted in log scale. (a) Clean speech without any processing. (b) Reverberated speech without any processing. (c) Clean speech with SSF processing. (d) Reverberated speech with SSF processing.

Onset enhancement and steady-state suppression by SSF processing reduced the difference between the processed powers of clean and reverberated speech than that between the unprocessed powers. However, the power contours of the processed signals for clean and reverberated speech are still different mainly in the boxes. The solid and dashed boxes represent the power contours corresponding to voiced speech with high power and unvoiced speech with powers concentrated in high-frequency channels, respectively. In the solid boxes, the power contours of reverberated voiced speech are more stretched than those of clean voiced speech, even after SSF processing is applied. On the other hand, useful features of the processed powers for reverberated speech in the dashed box were significantly removed by subtracting the low-pass-filtered powers of previous high-power voiced speech.

In applying (3), to avoid removing features that are useful for ASR in unvoiced speech such as fricatives, we estimate the probability that a frame corresponds to unvoiced speech with powers concentrated in high-frequency channels by measuring the *channel power ratio*, which is defined as the ratio of the power in high-frequency channels to the total power, given by

$$\zeta_c[m] = \frac{\sum_{l=l_u}^{L-1} \bar{P}[m,l]}{\sum_{l=0}^{L-1} \bar{P}[m,l]} \qquad (4)$$

where l_u determines the lowest channel index in high-frequency channels and $\bar{P}[m,l]$ describes the spectral power $P[m,l]$ with the reverberated components removed. Specifically, we calculate

$$\bar{P}[m,l] = \max\left(P[m,l] - \frac{1}{\alpha_{\max}} \sum_{\alpha=1}^{\alpha_{\max}} P[m-\alpha,l], \epsilon_g\right) \qquad (5)$$

where ϵ_g sets the floor value for $\bar{P}[m,l]$. This subtraction is performed because reverberated components that had high power in previous frames affect the power in the current speech frame, so the reverberated components need to be removed to allow the successful detection of unvoiced speech frames. For a frame

with a channel power ratio $\zeta_c[m]$, the subtraction amount in (3) is reduced so as to retain useful features by

$$\tilde{P}[m,l] = \max(P[m,l] - (1 - c_c\zeta_c[m])M[m,l], c_0 M[m,l]) \tag{6}$$

where c_c is a coefficient to adjust the dynamic range of $\zeta_c[m]$.

Additionally, differences between the power contours of processed signals for clean and reverberated voiced speech should be reduced to further improve the speech recognition performance. More aggressive suppression of reverberant components than that used in SSF processing can be considered; however, it may also remove many useful features, because it is very hard to estimate accurately the reverberant components. Therefore, instead of using aggressive suppression to obtain the processed power contours similar to those of clean speech, we boost the floor value in (6) to stretch the power contours along the time axis. In particular, the main difference happens in the reverberation of voiced speech with high power contours, and the voiced speech is detected based on harmonics. The *harmonic power ratio*, which is defined as the ratio of the power in harmonic-frequency bins to the total power, is introduced to measure a probability that a frame corresponds to voiced speech. Harmonic-frequency bins are the bins that represent integer multiples of the fundamental frequency.

Although many methods have been proposed to estimate the fundamental frequency (e.g., [11]–[13]), this letter employs a simple and effective autocorrelation-based method, in which the estimated fundamental frequency at frame m, $\hat{F}_0'[m]$, is obtained from the time-lag $\tau_0[m]$ that corresponds to the maximum autocorrelation, expressed as

$$\hat{F}_0'[m] = \frac{F_s}{\tau_0[m]}, \tag{7}$$

$$\tau_0[m] = \underset{\tau_{0,\min}<\tau<\tau_{0,\max}}{\arg\max} \sum_{k=0}^{N-1} X[m,k]X^*[m,k]e^{j2\pi k\tau/N} \tag{8}$$

where F_s denotes the sampling frequency, and $\tau_{0,\min} = \mathrm{round}(F_s/400\ \mathrm{Hz})$ and $\tau_{0,\max} = \mathrm{round}(F_s/70\ \mathrm{Hz})$ represent the time-lags corresponding to the maximum and minimum fundamental frequencies that normal speakers can utter, respectively. In practice, the value of $\hat{F}_0'[m]$ is averaged over adjacent frames to avoid abrupt changes as follows:

$$\hat{F}_0[m] = \frac{1}{2\beta_{\max}+1} \sum_{\beta=-\beta_{\max}}^{\beta_{\max}} \hat{F}_0'[m+\beta]. \tag{9}$$

To measure the harmonic power ratio for the current speech frame while excluding reverberant components from speech from previous frames, the power averaged over the previous $\alpha_{\max}$ frames is subtracted from the power at the current frame:

$$\bar{P}'[m,k] = \max\left(|X[m,k]|^2 - \frac{1}{\alpha_{\max}} \sum_{\alpha=1}^{\alpha_{\max}} |X[m-\alpha,k]|^2, \epsilon_f\right) \tag{10}$$

where ϵ_f sets the floor value for $\bar{P}'[m,k]$. Then, the harmonic power ratio estimated at frame m can be computed using

$$\zeta_h'[m] = \frac{\sum_{h=1}^{h_{\max}} \max_{\delta\in\Delta} \bar{P}'[m,\kappa(m,h)+\delta]}{\sum_{k=0}^{N/2} \bar{P}'[m,k]} \tag{11}$$

where $\kappa(m,h)$ denotes the frequency-bin index corresponding to the hth harmonic frequency of $\hat{F}_0[m]$, which is obtained using $\mathrm{round}(h \cdot \hat{F}_0[m] \cdot N/F_s)$; $h_{\max} = \mathrm{floor}(4\ \mathrm{kHz}/\hat{F}_0[m])$ is the number of harmonic frequencies in the band up to 4 kHz, which contains dominant harmonic components. Δ denotes the set of integer frequency-bin offsets from $-\delta_{\max}$ to $\delta_{\max}$ used to search harmonic components with inaccurate $\hat{F}_0[m]$, where $\delta_{\max}$ is set to $\mathrm{round}(70\ \mathrm{Hz} \cdot N/F_s)$. Similar to the estimation of $\hat{F}_0[m]$, $\zeta_h'[m]$ values are averaged over adjacent frames to obtain the desired harmonic power ratio as follows:

$$\zeta_h[m] = \frac{1}{2\beta_{\max}+1} \sum_{\beta=-\beta_{\max}}^{\beta_{\max}} \zeta_h'[m+\beta]. \tag{12}$$

For a frame with a $\zeta_h[m]$ value, the floor value is boosted up to the l_hth gammatone channel by

$$\tilde{P}[m,l] = \max(P[m,l] - (1 - c_c\zeta_c[m])M[m,l], c_s M[m,l]) \tag{13}$$

where

$$c_s = \begin{cases} \max(c_h\zeta_h[m], c_0), & l \le l_h, \\ c_0, & \text{otherwise.} \end{cases} \tag{14}$$

c_h is a coefficient to adjust the dynamic range of $\zeta_h[m]$.

Using the spectral reshaping approach described in [6] and [9], the channel weighting coefficient $w[m,l]$ is computed using

$$w[m,l] = \frac{\tilde{P}[m,l]}{P[m,l]}, \quad 0 \le l \le L-1. \tag{15}$$

Then, the spectral weighting coefficient $\mu[m,k]$ is obtained by

$$\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l[k]|}{\sum_{l=0}^{L-1} |H_l[k]|}, \quad 0 \le k \le N/2. \tag{16}$$

Assuming that the processed spectrum has the same phase as the original spectrum, the processed spectrum for the lower half of the frequency region is obtained using

$$\tilde{X}_e[m,k] = \mu[m,k]X_e[m,k], \quad 0 \le k \le N/2. \tag{17}$$

After invoking the Hermitian symmetry of the processed spectrum to obtain the remaining frequency components, the enhanced speech $\tilde{x}[n]$ is resynthesized using the inverse STFT and the overlap-add method as in [6] and [9].

III. Experimental Results

To evaluate SHARP processing as a preprocessing method for ASR, we conducted recognition experiments using the Wall Street Journal database and the Kaldi toolkit [14]. The recognition system was based on hidden Markov models (HMMs) with observation distributions of fully continuous Gaussian mixture models trained on 37 416 clean utterances (si284). The test set consisted of 836 utterances (dev93 and eval92). Speech recognition was based on the observed values of 13th-order mel-frequency cepstral coefficients with corresponding delta and acceleration coefficients. The cepstral coefficients were obtained from 23 mel-frequency bands with a frame size of 25 ms and a frame shift of 10 ms. We also compared our results using SHARP as described above to the improvements provided by the fMLLR method [8].

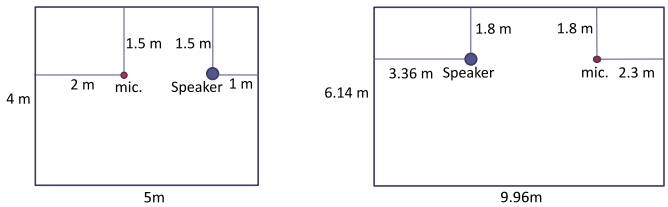

Fig. 3. Source and microphone positions to obtain reverberated speech as test data. Left panel: configuration for simulated speech. The room is 3 m high, and the source and microphones are 1.5 m above the floor. Right panel: configuration for live-recorded speech. The room is 2.47 m high, and the source and microphones are 1.3 m above the floor.

TABLE I
PARAMETER VALUES USED IN THE EXPERIMENTS

Parameter	N	L	λ	c_0	l_u	$\alpha_{\max}$	
Value	1,024	40	0.4	0.01	34	10	
Parameter	ϵ_g	c_c	F_s	$\beta_{\max}$	ϵ_f	c_h	l_h
Value	10^{-6}	0.1	16 000	1	10^{-6}	0.35	19

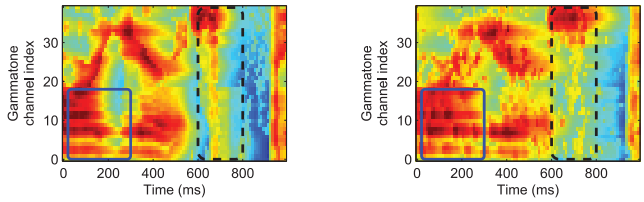

Fig. 4. Power spectra of clean and reverberated speech in gammatone channels processed using SHARP processing. The input clean and reverberated speech signals were the same as in Fig. 2. The values are depicted in log scale. (a) Clean speech with SHARP processing. (b) Reverberated speech with SHARP processing.

Test data with reverberated speech were obtained by convolving clean test data with room impulse responses generated by the image method (using the software package [15]), which simulates acoustics between two points in a rectangular room [16]. Fig. 3(a) depicts the configuration of the virtual room used to simulate the acoustic filters, which is the same virtual room configuration as in [9]. The reflection coefficient was selected to obtain a designated reverberation time RT_{60}.

Table I summarizes the parameter values used in the experiments. N, L, λ, and c_0 were set as recommended in [6]. $\alpha_{\max}$ was set to compute averaged powers over a time period of longer than 100 ms for appropriate smoothed power contours. $\beta_{\max}$ was chosen to avoid abrupt changes, and ϵ_g and ϵ_f were set to a small positive floor value. The values of l_u, l_h, c_c, and c_h were optimized empirically in pilot experiments.

Fig. 4 displays the power spectra of clean and reverberated speech using SHARP processing of (13). The input speech signals were the same as in Fig. 2. By subtracting the reduced amounts of low-pass-filtered power based on the channel power ratio and by boosting the floor value using the harmonic power ratio, the difference between the processed powers of clean and reverberated speech when using SHARP processing was much smaller than when using SSF processing, especially in the locations indicated by the two boxes.

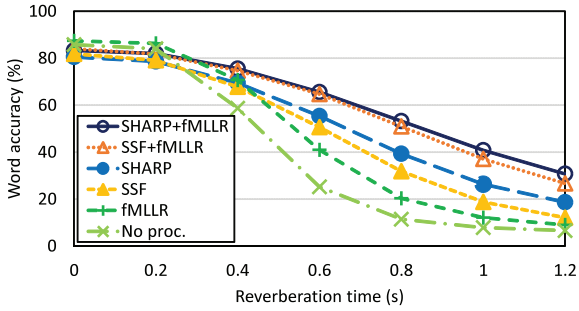

Fig. 5. Word accuracies obtained from SHARP processing.

TABLE II
WORD ACCURACIES (%) OBTAINED FOR LIVE RM DATA

Processing	No proc.	SSF	SHARP
W/o fMLLR	18.55	60.80	**70.56**
With fMLLR	37.41	81.80	**83.01**

Fig. 5 describes word accuracies obtained using either no processing, SSF processing, and SHARP processing. While both the SSF and SHARP methods achieve significant performance improvements in reverberant environments, SHARP processing provides greater recognition accuracies than SSF processing in highly reverberant environments and comparable accuracies in less reverberant environments. We also note that SHARP provides good results despite the use of a very simple autocorrelation-based method to estimate the fundamental frequency using (7)–(9) that is prone to estimating doubled or halved fundamental frequencies. While the use of fMLLR alone is less effective in highly reverberant environments, it does improve the performance of systems that already incorporate SHARP or SSF processing. The incorporation of fMLLR diminishes but does not eliminate the advantage of SHARP over SSF processing in highly reverberant environments. For environments with small reverberation times, the best performance is obtained with fMLLR alone. In general, the interaction of SHARP and fMLLR is complementary, as SHARP + fMLLR performs better than either alone on average.

To confirm the effectiveness of SHARP processing for real data, we repeated recognition experiments using the DARPA resource management (RM) database [17]. The acoustic models were based on the same types of HMMs trained on 3990 sentences from the original training set, and test data were obtained by rerecording the 300 test sentences in a normal office room using the configuration depicted in Fig. 3(b). Table II summarizes word accuracies for the live-recorded data, which are consistent with Fig. 5.

IV. CONCLUSION

In this letter, we present the SHARP preprocessing method, which extends the earlier SSF algorithm and makes use of stationary-component suppression using harmonics and the power ratio to achieve robust speech recognition. The SHARP method provides substantial improvements in recognition accuracy in highly reverberant environments compared to the earlier SSF algorithm. The use of fMLLR in reverberant environments is also beneficial but only if SHARP or SSF processing is also included.

REFERENCES

[1] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Hoboken, NJ, USA: Wiley, 2012.

[2] J. Droppo and A. Acero, "Environmental robustness," in *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. New York, NY, USA: Springer, 2008, pp. 653–680.

[3] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[4] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 3701–3704.

[5] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech recognition," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1053–1064, 1994.

[6] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Proc. INTERSPEECH Conf.*, Sep. 2010, pp. 2058–2061.

[7] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. Eur. Conf. Speech Commun. Technol.*, Sep. 1991, pp. 413–416.

[8] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[9] H.-M. Park, M. Maciejewski, C. Kim, and R. M. Stern, "Robust speech recognition in reverberant environments using subband-based steady-state monaural and binaural suppression," in *Proc. INTERSPEECH Conf.*, Sep. 2014, pp. 2715–2718.

[10] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica—Acta Acustica*, vol. 82, pp. 335–345, 1996.

[11] H. Quest, O. Schreiner, and M. R. Schroeder, "Robust pitch tracking in the car environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, pp. I-353–I-356.

[12] R. Petrick, K. Lohde, M. Lorenz, and R. Hoffmann, "A new feature analysis method for robust ASR in reverberant environments based on the harmonic structure of speech," in *Proc. Eur. Signal Process. Conf.*, Aug. 2008, pp. 1–5.

[13] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3690–3700, 2004.

[14] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recog. Understand.*, Dec. 2011.

[15] S. G. McGovern, "Room impulse response generator," in *MATLAB Central File Exchange* [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator, Jan. 2013.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[17] P. Price, W. M. Fisher, J. Bernstein, and D. Pallet, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1988, pp. 651–654.