Acoustic Analysis of Whispered Speech for Phoneme and Speaker Dependency

Xing Fan, Keith W. Godin, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, U.S.A.
xxf064000@utdallas.edu, godin@ieee.org, john.hansen@utdallas.edu

Abstract

Whisper is used by speakers in certain circumstances to protect personal information. Due to the differences in production mechanisms between neutral and whispered speech, there are considerable differences between the spectral structure of neutral and whispered speech, such as formant shifts and shifts in spectral slope. This study analyzes the dependency of these differences on speakers and phonemes by applying a Vector Taylor Series (VTS) approximation to a model of the transformation of neutral speech into whispered speech, and estimating the parameters of this model using an Expectation Maximization (EM) algorithm. The results from this study shed light on the speaker and phoneme dependency of the shifts of neutral to whisper speech, and suggest that similarly derived model adaptation or compensation schemes for whisper speech/speaker recognition will be highly speaker dependent.

Index Terms: whispered speech, speech analysis

1. Introduction

Whispered speech is a natural mode of speech production, employed in public situations in order to protect personal information. A customer might whisper to provide information regarding their date of birth, credit card information, and billing address in order to make hotel, flight, or car reservations by telephone, or a doctor might whisper in order to discuss patient medical records in public. Studies including [1] and [2] have demonstrated the detrimental effects of whisper on speech recognition and speaker recognition systems.

Whispered speech is defined as the absence of periodic vocal fold vibration in the production of phonemes that otherwise include such vibration [3], and the acoustic properties of whispered speech are the subject of ongoing study. Relative to their neutral speech counterparts, whispered phones undergo formant shifts, expansion of formant bandwidths, and shifts in cepstral distances [1, 3, 4]. However, study of the acoustic properties of whispered phones has not generally examined individual speaker differences in variations of these parameters. Such an analysis is crucial to understanding both the whisper production process and the design of robust speech systems. It has been shown, for example, that such individual production differences result in great variation in the performance of speaker recognition systems [5]. In that study, it was observed that recognition of the identity of some speakers degraded significantly when recognition was performed on whispered speech, while the recognition of other speakers was relatively unaffected by whisper.

Performance of speech systems on whisper may be improved with the incorporation of whispered speech from target speakers to be used in model adaptation or in the development of feature compensation schemes [1]. However, some studies have considered feature and model compensation schemes that work in the absence of such data, in order to address those real-world situations in which whisper data from target speakers is not feasible to collect [5]. The design of such schemes relies on what is known generally about whispered speech, and so it is important for the development of these systems to continue the study of the spectral and time domain differences between whispered and neutral speech.

This study compares the smoothed spectral envelope of whispered and neutral speech by applying a time domain linear model, and examines the differences in estimated model parameters between speakers and phonemes. The research questions for the present study are: are differences between whispered and neutral speech consistent among all speakers? If they are not consistent across speakers, are they more consistent when controlling for vowel/consonant character of an utterance? The answers to these questions could have implications for the design of speech systems, and could suggest directions for further evaluation of the acoustic nature of whispered speech. If spectral differences between whisper and neutral are consistent across speakers, a feature or model transformation estimated using whispered adaptation data from several non-target speakers could be applied to all whispered enrollment and test data. On the other hand, if spectral differences between whispered and neutral speech are inconsistent across speakers, system designs must explore adaptation methods that could estimate the particulars of a given enrollment speaker's whispered speech.

In order to compare the spectral structure of whispered and neutral speech, a time domain linear model plus noise model is motivated, and transformed into the cepstral domain. The parameters of the cepstral domain model are approximated using Vector Taylor Series (VTS) and the parameters of the final, approximate model are estimated by applying the Expectation-Maximization (EM) algorithm. The following sections of this paper discuss the whispered/neutral corpus used in this study, the neutral-whisper transformation model applied in this study, the experimental procedure, results, and conclusions.

2. Corpus

The corpus developed in [4] supplies the whispered/neutral paired utterances used in this study. Ten male native speakers of American English were recruited to speak 10 TIMIT sentences in both whisper and neutral. In this way, the same phoneme context is provided across speakers and speech mode. Recordings occurred in an ASHA-certified single-walled soundbooth, using

This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067 (Approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

a Shure Beta 53 head-worn close talking microphone, and were digitized and recorded using a Fostex D824 digital recorder at 44.1kHz, with 16 bits per sample, and downsampled to 16kHz for this study.

Additionally, the TIMIT corpus [6] is used to develop initial models for a neutral speech recognition system.

3. Speech transformation model

This study models the transformation of neutral speech $\mathbf{ne}(\mathrm{t})$ into whisper $\mathbf{wh}(\mathrm{t})$ using a linear time-invariant (LTI) h(t) plus a noise term n(t). This assumption serves as a first order approximation; the model is limited in its power to capture the formant shifts of whispered speech, but will capture aspects of the smoothed spectral envelope of whispered speech. The smoothed spectral envelope is used to represent the spectral information in the front-end processing for most state of the art speech systems. Due to the introduction of convolution, the complexity of the parameters to be estimated decreases significantly compared with other cepstral domain linear regression thus reducing the chance of overfitting and resulting an estimation close to the ground truth.

$$\mathbf{wh}(\mathrm{t}) = \mathbf{ne}(\mathrm{t}) * \mathrm{h}(\mathrm{t}) + \mathrm{n}(\mathrm{t}). \tag{1}$$

This model and the accompanying Vector Taylor Series (VTS) estimation method described in Section 4.1 have been used to improve the robustness of speech recognition systems in noisy environments [7] [8]. The model estimation process is performed in the cepstral domain and is based on estimating a transform of the cepstral distributions estimated from neutral-speech acoustic models.

For simplicity, it is assumed that the phase of neutral and whispered speech is synchronized. Hence, in the MFCC domain, the relationship between whisper and neutral is:

$$\mathbf{wh} = \mathbf{ne} + h + g(\mathbf{ne}, h, n), \tag{2}$$

$$g(\mathbf{ne}, h, n) = C\log(1 + \exp(C^{-1}(n - \mathbf{ne} - h))) \tag{3}$$

where C^{-1} is the pseudo-inverse DCT matrix. The whisper noise distortion n is assumed Gaussian distributed with zero mean μ_n and a diagonal covariance matrix Σ_n. The filter h is assumed to be a fixed vector with deterministic values that represents the shape of the smoothed spectral envelop of h(t). Applying the first order VTS approximation around the point $(\mu_{\mathbf{ne}}, \mu_h, \mu_n)$, we have

$$\begin{aligned}\mathbf{wh} &\approx \mu_{\mathbf{ne}} + \mu_h + g(\mu_{\mathbf{ne}}, \mu_h, \mu_n)\\ &\quad + G(\mathbf{ne} - \mu_{\mathbf{ne}}) + G(h - \mu_h) + F(n - \mu_n),\end{aligned} \tag{4}$$

where,

$$\begin{aligned}&\frac{\partial \mathbf{wh}}{\partial \mathbf{ne}}|_{\mu_{\mathbf{ne}},\mu_h,\mu_n} = \frac{\partial \mathbf{wh}}{\partial h}|_{\mu_{\mathbf{ne}},\mu_h,\mu_n} = G\\ &\frac{\partial \mathbf{wh}}{\partial n}|_{\mu_{\mathbf{ne}},\mu_h,\mu_n} = I - G = F\\ &G = C \cdot diag\{\frac{1}{1 + \exp(C^{-1}(\mu_n - \mu_{\mathbf{ne}} - \mu_h))}\} \cdot C^{-1},\end{aligned} \tag{5}$$

where $diag\{\}$ stands for a diagonal matrix with its diagonal component value equal to the value of the vector in the argument. Taking the expectation and variance operations of both sides of Eq (4), the resulting static $\mu_{\mathbf{wh}}$ and $\Sigma_{\mathbf{wh}}$ are (noting that the filter h is a fixed vector):

$$\begin{aligned}\mu_{\mathbf{wh}} &\approx \mu_{\mathbf{ne}} + \mu_h + g(\mu_{\mathbf{ne}}, \mu_h, \mu_n)\\ \Sigma_{\mathbf{wh}} &\approx G\Sigma_{\mathbf{ne}}G^t + F\Sigma_n F^t\end{aligned} \tag{6}$$

A similar procedure is applied to estimate the parameters for delta and delta/delta portions of MFCC features.

Conceptually, μ_h in Eq. 6 will serve in this study as the estimate of the effect of whisper production on the speech acoustic signal, and will be compared across speakers and phonemes.

4. Experimental method

To estimate μ_h from Eq. 6 for the comparisons made in this study, a Hidden Markov Model (HMM) based Automatic Speech Recognition (ASR) system for neutral speech is developed, and a transformation based on Eq. 6 of the model parameters of that system is estimated using Vector Taylor Series (VTS).

The neutral ASR system is developed using the TIMIT corpus [6]. Speech is windowed with a Hamming window of 25ms, with a 10ms overlap. 13-dimensional MFCCs, appended with their first- and second-order time derivatives are used as acoustic features. Each HMM is left-to-right with 3 states, with 16 Gaussian mixtures per state. Finally, using the 10 sentences of neutral speech from each target speaker in the corpus, neutral HMMs for each target speaker are adapted from the TIMIT HMMs using Maximum Likelihood Linear Regression (MLLR).

The word accuracy of the neutral-adapted ASR system on neutral test data is 83.44%. However, the word accuracy drops to 22.62% when tested with whispered speech, demonstrating the degradation in performance due to whisper/neutral mismatched train/test conditions.

From the neutral speaker adapted HMMs, $\mu_{\mathbf{ne}}$ and $\Sigma_{\mathbf{ne}}$ in Eq. 6 are simply the mean and covariance of each mixture Gaussian in every state. Starting from these, the next two subsections describe how μ_{wh} and Σ_{wh} may be estimated using the 10 sentences of whisper adaptation data for each speaker. The final transform computed in the process of this estimation, μ_h from Eq. 6, is the focus of the transform analysis in this study.

4.1. Estimation of filter parameter and noise

The Expectation-Maximization (EM) algorithm is applied to estimate μ_h. Given a whispered utterance $\mathbf{wh}$, the EM auxiliary function is:

$$Q(\lambda|\bar{\lambda}) = \sum_t \sum_{s,m} \gamma_{tsm}^{phone} \log(p(\mathbf{wh}_t|s, m, \lambda_{phone})) \tag{7}$$

where $p(\mathbf{wh}_t|s, m, \lambda_{phone}) \sim N(\mathbf{wh}_t; \mu_{\mathbf{ne},sm}, \Sigma_{\mathbf{ne},sm})$ and γ_{tsm}^{phone} is the posterior probability of the m^{th} Gaussian pdf in the s^{th} state of HMM corresponding to the specific "*phone*" for the t^{th} frame in $\mathbf{wh}$. γ_{tsm}^{phone} can be calculated using the forward-backward algorithm. Instead of updating the parameters of HMMs [7], [8], the h_{phone} is estimated for each particular phone in the utterances. This supports exploration of the differences between whispered and neutral speech at the phone level. To ensure the speaker dependent HMMs fit to the adaptation data, only phonemes with sufficient data for MLLR are considered.

In the M-step, we take the derivatives of Q with respect to μ_h^{phone}. The update formula for each μ_h^{phone} is found by setting the derivatives to zero:

$$\begin{aligned}&\mu_h^{phone} = \mu_{h,0}^{phone} + \{\sum_{t=start}^{end} \sum_{s,m} \gamma_{tsm}^{phone} G_{s,m}^t \Sigma_{wh,sm}^{-1} G_{s,m}\}^{-1}\\ &\{\sum_{t=start}^{end} \sum_{s,m} \gamma_{tsm}^{phone} G_{s,m}^t \Sigma_{wh,sm}^{-1}\\ &[wh_t - \mu_{ne,sm} - \mu_{h,0}^{phone} - g(\mu_{ne,sm}, \mu_{ne,sm}, \mu_n)]\}\end{aligned} \tag{8}$$

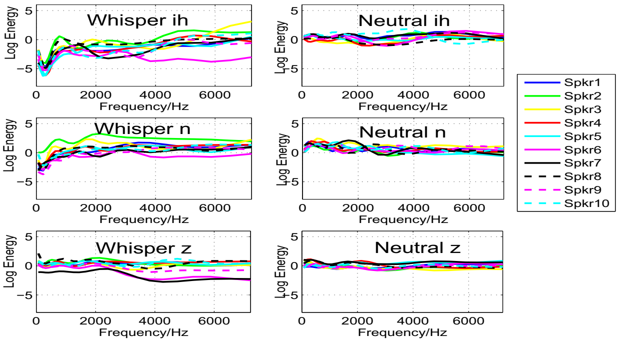

Figure 1: *Examples of the average estimated μ_H.*

The whisper noise n is assumed stationary, thus $\mu_{\Delta n} = 0$ and $\mu_{\Delta\Delta n} = 0$. The Σ_n^{phone} is updated as in [7] and [8] using Newton's method:

$$\Sigma_n^{phone} = \Sigma_{n,0}^{phone} - [(\frac{\partial^2 Q}{\partial^2 \Sigma_n^{phone}})^{-1}(\frac{\partial Q}{\partial \Sigma_n^{phone}})] \qquad (9)$$

For $\Sigma_{\Delta n}$ and $\Sigma_{\Delta\Delta n}$, a similarly derived update formula is employed.

4.2. Algorithm implementation

Given the MFCC feature vectors of a phone in whispered speech, the procedure for estimating the h and n parameters in Eq. 3 are:

1. Obtain the corresponding $\mu_{\mathbf{ne},sm}^{phone}$ and $\Sigma_{\mathbf{ne},sm}^{phone}$ from the trained monophone neutral HMMs. Set the initialized μ_h^{phone} to zero.
2. Compute the G_{sm} and F_{sm} corresponding to each original neutral Gaussian pdf in the phone HMM with Eq. (5) by using the current μ_h^{phone}. Update each of the phone HMM parameters with Eq. (6) and compute the posterior probabilityγ_{tsm}^{phone} given the whispered phone.
3. Update the μ_h^{phone}, Σ_n^{phone}, $\Sigma_{\Delta n}$ and $\Sigma_{\Delta\Delta n}$ using Eq. (8, 9).
4. Decode the whispered phoneme with the updated HMMs and compute the likelihood. If the likelihood converges, record the μ_h^{phone}. Otherwise, repeat the process by going back to Step 2.

5. Results

The parameter μ_h from Eq. 6 is estimated for each whispered phone in the corpus. To ensure that the speaker-dependent HMMs fit well to the neutral adaptation data, only phones of sufficient length for MLLR are included in the analysis. The μ_h that results from the preceding algorithm is in the cepstral domain; for the following analysis, it is converted to the frequency domain by applying C^{-1}, the pseudo inverse of the DCT matrix.

Fig. 1 shows the average μ_H of the whispered phonemes /ih/, /n/, and /z/ for each speaker. For comparison, the average μ_H of the neutral phonemes is provided. The figure shows that for vowels and nasals, the transformation of neutral speech into whisper includes compression of the energy of the neutral speech especially in the lower frequency range of about 0-2kHz. For the voiced fricative "z", the transform is near zero, implying that the neutral speech and the whispered speech are very similar. This is also the case for the neutral phones. This confirms the consistency of our estimation method and its implementation. The next sections divide the μ_h into vowels and consonants for separate analysis.

5.1. Results of the vowel analysis

The following analysis arbitrarily divides the frequency range from 0-8kHz kHz into 3 subbands: S1(0-2700 Hz), S2(2700-4000 Hz), and S3(4000-8000 Hz), representing approximately a phone dependent frequency range, a speaker dependent frequency range, and the remaining high frequency range. The subbands μ_h^{S1}, μ_h^{S2}, and μ_h^{S3} can be obtained from μ_h through simple linear algebra. Fisher's discriminant power is used to analyze separately the dependence of each subband on inter-speaker variation and inter-phone variation.

A greater magnitude of the discriminant power implies better separation between the given clusters in the sample space. Given K classes of μ_h^{S1} that constitute the sample space Φ_K, there are K cluster means $\mu_{h,k}^{S1}$ and K cluster variances $\Sigma_{h,k}^{S1}$, where $1 \leq k \leq K$. The mean of the cluster means $\mu_{h,k}^{S1}$ is denoted $\bar{\mu_g}$. Assuming there are W_k samples in each class of $\mu_{h,k}^{S1}$, Fisher's discrimination power is:

$$F(\mu_h^{S1}) = \frac{S_B}{S_w} = \frac{\| \sum_{k=1}^{K} W_k(\mu_{h,k}^{S1} - \bar{\mu_g})(\mu_{h,k}^{S1} - \bar{\mu_g})^T \|_2^2}{\sum_{k=1}^{K} \Sigma_{h,k}^{S1}} \qquad (10)$$

The Fisher discrimination power $F_{p,s}^{S1}$ is computed for each speaker s, treating each phoneme p as a class. This measures the inter-phoneme variability in subband $S1$ of μ_h^{S1} for speaker s. F_p^{S1} denotes the mean of this quantity across all speakers. The Fisher discrimination power $F_{s,p}^{S1}$ is computed for each phoneme p, treating each speaker s as a class. This measures the inter-speaker variability in subband $S1$ of μ_h^{S1} for each phoneme. F_s^{S1} denotes the mean of this quantity across all phonemes. The quantities F_p^{S2}, F_p^{S3}, F_s^{S2}, and F_s^{S3} are computed similarly.

Table 1: *Fisher's discrimination power in discriminating vowels within each speaker, and discriminating speakers within each vowel, for subbands $S1$, $S2$ and $S3$.*

Subband	F_p^{Sx}	F_s^{Sx}
$S1$	2.46	1.94
$S2$	2.32	3.69
$S3$	2.79	6.65

Table 1 shows the results. The table shows that the F_s^{Sx} increases with increasing frequency subband, while F_p^{Sx} stays similarly with relatively small value. This suggests that the differences between whispered vowels and neutral vowels are similar across the frequency range given a specific speaker. In low frequency domain, the difference also shares similarity among all speakers with slightly changes given different phonemes. With the increase of frequency, the phoneme-dependency is lost while the speaker-dependency difference strongly remains. Fig. 2 demonstrates this. On the figure are the μ_h for each speaker in the corpus for the vowels /ax/ and /ih/. The figure shows that the μ_h are generally speaker dependent especially in higher frequencies while varies slightly across vowels in lower frequencies.

Further analysis focuses on subbands $S1$ and $S2$ in order to investigate the dependency of μ_h on speakers for these more phone dependent subbands, as S3 appears to be highly speaker dependent. The standard deviation (std) of μ_h^{S1-S2} is calculated for each vowel to measure the speaker diversity. Each vowel is compared on its relative height in American English. Comparing the relative tongue heights with the std of μ_h^{S1-S2} across the

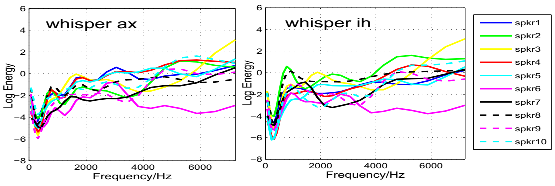

Figure 2: *μ_h for each speaker for each of the vowels /ax/ and /ih/.*

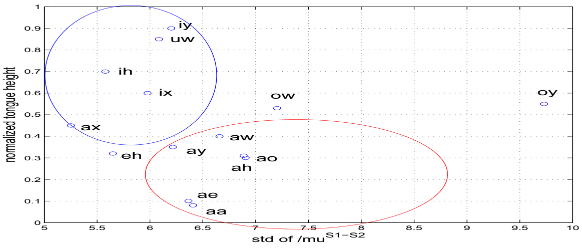

Figure 3: *Tongue height vs. std of μ_h^{S1-S2}.*

speakers suggests a possible relationship between the std and the relative tongue height of vowels. Fig. 3 shows that whisper vowels with higher relative tongue height are clustere in a region with lower inter-speaker variability of μ_h^{S1-S2}. Vowels with lower relative tongue height appear more to have neutral to whisper transformations μ_h^{S1-S2} of greater speaker diversity.

5.2. Results of the consonant analysis

For the following analysis, the consonants are grouped into five categories: 1. Unvoiced consonants (UVC), which include unvoiced stops, affricates, and fricatives, 2. Voiced consonants from stops, affricates, and fricatives (VC) that can be mapped to the unvoiced consonants, 3. Nasals, 4. Liquids and 5. Semivowels. In order to analyze the impact of the absence of voiced excitation on consonants, the spectral tilt of the neutral-whisper transfer function μ_H is measured by using a first order linear interpolation of $\mu_H = C_1[log\,frequency] + C_2$. Fig. 3(a) shows that UVC and VC share a slight spectral tilt change in the smoothed spectral envelope, despite the absence of voiced excitation in whispered VC. However, the spectral tilt of nasals, semi-vowels, and liquids undergoes greater change from neutral to whisper than the UVC and VC. This suggests that nasals, semi-vowels, and liquids undergo greater change in the spectral domain due to whispering.

In order to investigate the speaker and phoneme dependency of μ_h for consonants, $F_s(\mu_h^{S1,S2,S3})$ and $F_p(\mu_h^{S1,S2,S3})$ are calculated in the same way as for the analysis of the vowels. Considering the similarity between whispered speech and neutral speech in the production of stops, fricatives and affricates, only liquids, semi-vowels, and nasals are considered in this part of the experiment. Table 2 shows μ_h is highly phoneme dependent in the lower frequency range, which confirms the observation of $C1$ and $C2$. μ_h becomes more speaker dependent with increasing frequency.

6. Conclusions

This study applied a VTS approximation to a transformation model of neutral to whispered speech in order to compare neutral and whispered speech, and in order to analyze the dependency of the difference on phonemes and speakers.

The experimental results suggest that for vowels, the difference between whispered and neutral speech is generally consistent across speakers, especially beyond 4kHz. The results also suggest that the differences between whispered and neutral speech are also more consistent across the vowels for frequencies below 4kHz than above. A possible relationship between average tongue height of vowels and speaker diversity in vowel production was observed. Shifts in spectral tilt due to whisper of consonants were shown to differ across five consonant categories. Finally, the results suggest that the spectral differences due to whisper vary across liquids, semivowels, and nasals.

Table 2: *Fisher's discrimination power in discriminating consonants within each speaker, and discriminating speakers within each consonants, for subband S1, S2 and S3.*

Subband	F_p^{Sx}	F_s^{Sx}
S1	5.36	2.13
S2	2.97	2.74
S3	2.51	3.82

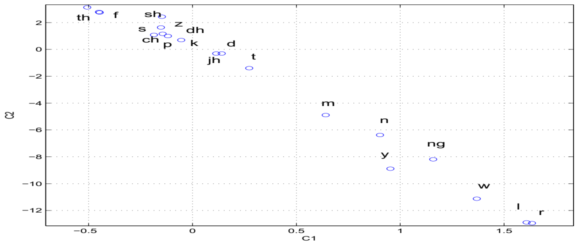

Figure 4: *Distribution of $C1$ and $C2$ for consonants.*

7. References

[1] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Comm.*, vol. 45, pp. 139–152, 2005.

[2] Q. Jin, S. S. Jou, and T. Schultz, "Whispering speaker identification," in *IEEE Intl. Conf. on Multimedia and Expo*, (Beijing, China), pp. 1027–1030, Jul. 2007.

[3] S. T. Jovicic, "Formant feature differences between whispered and voiced sustained vowels," *Acustica-acta*, vol. 84, pp. 739–743, 1998.

[4] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: whisper through shouted," in *INTERSPEECH 2007*, (Antwerp, Belgium), pp. 2289–2292, Aug. 2007.

[5] X. Fan and J. H. L. Hansen, "Acoustic analysis for speaker identification of whispered speech," in *IEEE Intl. Conf. Acoustics, Speech, and Sig. Proc. (ICASSP)*, (Dallas, U.S.A.), pp. 5046–5049, Mar. 2010.

[6] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium*, (Philadelphia, U.S.A.), 1993.

[7] J. Li, D. Yu, L. Deng, Y. Gong, and A. Acero, "A unified framework of hmm adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Lang.*, vol. 23, pp. 389–405, 2009.

[8] O. Kalinli, M. L. Seltzer, and A. Acero, "Noise adaptive training using a vector taylor series approach for robust automatic speech recognition," in *IEEE Intl. Conf. Acoustics, Speech, and Signal Proc. (ICASSP)*, (Taipei, Taiwan), pp. 3825 – 3828, Apr. 2009.