SPEAKER VERIFICATION BASED PROCESSING FOR ROBUST ASR IN CO-CHANNEL SPEECH SCENARIOS

Seyed Omid Sadjadi[1*] *and Larry P. Heck*[2]

[1]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas
omid.sadjadi@ieee.org

[2]Microsoft Research
larry.heck@ieee.org

ABSTRACT

Co-channel speech, which occurs in monaural audio recordings of two or more overlapping talkers, poses a great challenge for automatic speech applications. Automatic speech recognition (ASR) performance, in particular, has been shown to degrade significantly in the presence of a competing talker. In this paper, assuming a known target talker scenario, we present two different masking strategies based on speaker verification to alleviate the impact of the competing talker (a.k.a. masker) interference on ASR performance. In the first approach, frame-level speaker verification likelihoods are used as reliability measures that control the degree to which each frame contributes to the Viterbi search, while in the second approach time-frequency (T-F) level speaker verification scores form soft masks for speech separation. Effectiveness of the two strategies, both individually and in combination, are evaluated in the context of ASR tasks with speech mixtures at various signal-to-interference ratios (SIR), ranging from 6 dB to -9 dB. Experimental results indicate efficacy of the proposed speaker verification based solutions in mitigating the impact of the competing talker interference on ASR performance. Combination of the two masking techniques yields reductions as large as 43% in word error rate.

Index Terms— ASR, co-channel speech, soft masking, speaker verification

1. INTRODUCTION

With the proliferation of mobile devices which provide hands-free voice-enabled applications, there is a growing need to design algorithms that can improve robustness of automatic speech systems against various noise sources that are typically active at the same time. Co-channel speech [1, 2, 3, 4, 5, 6], for example, that occurs in monaural recordings of two or more simultaneous talkers, can cause significant degradations in performance of automatic speech recognition (ASR) systems. Co-channel speech is common in hands-free speech applications such as voice interaction with in-vehicle infotainment systems and gaming consoles as well as information and services access using a speaker-phone.

Unlike human listeners who possess a remarkable, yet seemingly simple, ability to segregate and subsequently attend to a specific talker (or in general a single sound source) within an acoustic mixture [7, 8, 9], machine listeners have been shown to face a great challenge while dealing with such scenarios [10]. This has motivated extensive research effort in the past several decades to understand the human auditory perception mechanism and design algorithmic solutions that can mimic this mechanism [11, 12, 13, 14, 15]. In addition to the computational auditory scene analysis (CASA) solutions [16, 17, 18, 19], which are bottom-up methods inspired by psychoacoustics principles, there have been several model-based top-down techniques proposed in the literature to tackle the problem of ASR in co-channel speech scenarios [6, 20, 21, 22, 23]. It is worth noting that traditional single-channel speech enhancement techniques (see [24] for a review) are not applicable for suppressing the competing talker in co-channel speech because the interference itself is speech and its time-varying statistics are neither known *a priori*, nor can they be estimated from the background.

In this paper, assuming a known target talker scenario, we present two different masking solutions based on speaker verification to cope with co-channel speech for ASR. In the first solution, normalized frame-level speaker verification likelihoods are used as reliability measures; if a frame is likely to have been produced by the target talker, it is viewed as reliable and contributes fully to the Viterbi search otherwise it is considered as unreliable (i.e., masked) and its contribution to the search is discounted. This method can be categorized as a missing-feature technique [25] which operates at frame-level (as opposed to time-frequency level). In the second solution, speaker verification scores are computed at time-frequency (T-F) level to serve as a soft mask for reconstruction-based speech separation [19, 20, 21, 22, 23]. The reconstructed speech signal for each target talker is then passed to the ASR system for transcription. This technique is similar to CASA-based speech separation approaches [17, 18] in the segmentation stage, however instead of using grouping cues [15] such as common periodicity, common onset/offset, and common amplitude modulation, the verification likelihoods are used to determine which regions of the auditory spectrogram belong to the target signal. In other words, grouping cues are learned using generative Gaussian mixture models (GMM) from training data for each target speaker. We employ the GMM-UBM framework [26] to compute speaker verification scores for both solutions. Effectiveness of the proposed strategies are evaluated, both individually and in combination, in the context of ASR tasks using audio material from the speech separation challenge (SSC) [27]. We use the hidden Markov model toolkit (HTK [28]) in our ASR experiments.

2. SPEAKER VERIFICATION BASED MASKING

In this section we provide descriptions of the two speaker verification based masking strategies for robust ASR in co-channel speech scenarios. Our goal is to suppress the competing talker interference

*This work was done while Omid Sadjadi was a Research Intern at Microsoft.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

either in the ASR decoding or indirectly at the signal level through speech separation. Unless stated otherwise, our assumption is that the target talker in the mixture is known.

2.1. Time-Level Soft Masking

We first formulate the time-level soft masking method as the joint maximization of the *a posteriori* probability of the word sequence and the target talker given the observed acoustic mixture, and then show that the resulting solution can be implemented as a frame-level soft masking in the Viterbi search.

Expressed mathematically, the goal here is to find the word sequence from the target talker such that the joint probability among all possible word sequences W and talkers S, conditioned on observations O, is maximized. The observations can take many forms, such as sequence of short-term cepstral feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ such as mel-frequency cepstral coefficients (MFCC) and power-normalized cepstral coefficients (PNCC) [29], or sequence of longer term prosodic feature vectors $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_Q\}$ [30, 31]. Given this set of observations $O = \{\mathbf{X}, \mathbf{F}\}$, we can define the joint maximization problem as,

$$\begin{aligned}
\{\hat{W}, \hat{S}\} &= \arg\max_{W,S} \; p(W, S \mid O) \\
&= \arg\max_{W,S} \; p(\mathbf{X}, \mathbf{F} \mid W, S) \cdot p(W, S) \\
&= \arg\max_{W,S} \; p(\mathbf{X} \mid \mathbf{F}, W, S) \cdot p(\mathbf{F} \mid W, S) \cdot p(W, S) \\
&= \arg\max_{W,S} \; p(\mathbf{X} \mid \mathbf{F}, W, S) \cdot p(\mathbf{F} \mid W, S) \cdot p(W \mid S) \\
&\qquad \cdot p(S) \\
&= \arg\max_{W,S} \; \underbrace{p(\mathbf{X} \mid W)}_{SI-Speech} \cdot \underbrace{\frac{p(\mathbf{X} \mid \mathbf{F}, W, S)}{p(\mathbf{X} \mid W)}}_{Speaker} \cdot \underbrace{p(\mathbf{F} \mid W, S)}_{Prosody} \\
&\qquad \cdot \underbrace{p(W \mid S)}_{SD-LM} \cdot \underbrace{p(S)}_{Prior}. \qquad (1)
\end{aligned}$$

These five components represent separate knowledge sources, including a speaker-independent recognizer (SI-Speech), a normalized speaker recognizer (Speaker), a prosodic subsystem (Prosody), a speaker-dependent language model (SD-LM), and a prior for the given speaker. The knowledge sources can be combined at different levels, i.e., from time-frequency (T-F) level to frame-level to utterance-level, although their availability is dependent on the task and application (see [5] for more details). For the ASR task considered in this paper, the use of prosody is not explored and all speakers are given the same prior. In addition, all talkers follow the same fixed grammar rule in their speech, therefore the speaker-dependent language model becomes uninformative. Accordingly, the maximization problem for co-channel speech recognition can be simplified as,

$$\{\hat{W}, \hat{S}\} = \arg\max_{W,S} \; p(\mathbf{X} \mid W) \cdot p(W) \cdot \frac{p(\mathbf{X} \mid W, S)}{p(\mathbf{X} \mid W)}. \qquad (2)$$

The terms on the right hand side of (2) represent speech recognition and speaker verification scores which can be combined in the search at various resolutions (from frame to utterance level). The combination at the frame-level, which is considered in this paper, can be accomplished in the forward pass of the Viterbi search. The one-pass approach is practical for applications with small number of speakers such as co-channel speech separation.

To combine the ASR and speaker verification systems for recognition of co-channel speech, the verification score is treated as a reliability measure. If a frame is likely to have been produced by the target speaker, it is viewed as reliable and contributes fully to the search, otherwise its contribution is discounted. More precisely, we use the verification likelihoods to weight the frame likelihoods from the ASR system as,

$$\Lambda_{total}(\mathbf{x}_t) = \Phi(\Lambda_{speaker}(\mathbf{x}_t)) \cdot \Lambda_{speech}(\mathbf{x}_t), \qquad (3)$$

where Φ is a sigmoid function that maps the verification likelihoods into [0 1] range, which is defined as,

$$\Phi(y) = \frac{1}{1 + \exp\big(-b(y - \theta)\big)}. \qquad (4)$$

In this paper, the verification likelihood for frame t, $\Lambda_{speaker}(\mathbf{x}_t)$, is computed within a text-independent GMM-UBM framework as,

$$\Lambda_{speaker}(\mathbf{x}_t) = \frac{1}{M} \sum_{i=t-\frac{M}{2}}^{t+\frac{M}{2}} \log p(\mathbf{x}_i \mid \lambda_{tgt}) - \log p(\mathbf{x}_i \mid \lambda_{imp}), \qquad (5)$$

where M is the length of a sliding window over which the smoothed verification score is calculated. Here, λ_{tgt} and λ_{imp} denote the speaker models for the target and interfering (or impostor) talkers, respectively. The availability of *a priori* knowledge regarding the target and impostor talkers is dependent on the application. Often, it is not unrealistic to assume that the target talker is known while the interfering talker is not, in which case a speaker-independent (SI) background model is used to represent the impostors.

The mapped verification likelihoods in (3) can be viewed as soft masks that suppress the competing talker's speech at the decoding stage of ASR. Accordingly, this method can be classified as a frame-level (as opposed to time-frequency level) missing-feature technique.

2.2. Time-Frequency Level Soft Masking

Time-frequency (T-F) masking techniques, which are essentially based on psychoacoustic principles observed in the human auditory perception mechanism, have been widely applied to improve speech intelligibility for both man and machine listeners in the presence of interfering maskers [11, 15, 17, 24, 25]. The T-F mask, which can be in binary (i.e., 0 or 1) or soft (e.g., in [0 1] range) form, is either used to segregate and reconstruct a foreground stream corresponding to the target signal within the acoustic mixture, or employed as a reliability measure for different regions in the auditory spectrogram of the mixture for missing-feature speech recognition. In either case, the T-F mask estimation primarily involves two stages: segmentation and grouping. In the segmentation stage, the mixture is decomposed into T-F units that represent the signal at a specific time frame and frequency channel, while in the grouping stage, the T-F units are merged according to psychoacoustic cues such as common periodicity, common amplitude modulation, common onset/offset, and temporal continuity.

In this paper, however, we take a different grouping approach based on speaker verification to estimate the soft T-F mask for co-channel speech separation. A block diagram illustrating the proposed soft mask estimation approach is shown in Fig. 1. After segmenting the speech mixture into T-F units using a 64-channel gammatone filterbank, we extract the amplitude modulation spectrogram (AMS) features from 32 ms frames at a 100 Hz rate [32, 33] to parametrize the signal in each T-F unit. Similar to the time level soft masking

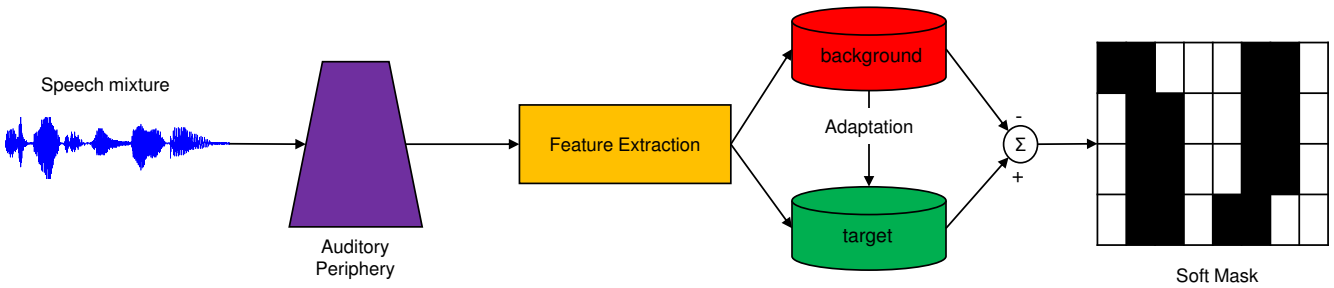

Fig. 1. Block diagram of the proposed speaker verification based technique for time-frequency soft mask estimation in co-channel speech.

technique described in the previous section, a GMM-UBM framework is then used to compute the speaker verification likelihoods at the T-F level according to (5), assuming that the target talker is known a priori in the mixture. Next, the T-F level verification likelihoods are smoothed across both time frames and frequency bands using 7 and 3 point moving average filtering, respectively. Finally, the smoothed likelihoods are mapped into [0 1] range using the sigmoid function, Φ, defined in (4), and the target talker's speech reconstructed with the estimated soft mask using the method described in [33]. Computing the T-F mask with this approach obviates the need for reliable multi-pitch tracking in voiced regions as well as onset/offset estimation in unvoiced segments. Here, grouping cues are learned using generative GMMs from training data for each speaker and the verification likelihoods determine which regions of the auditory spectrogram belong to the target talker.

Fig. 2 shows sample spectrograms for the the sentence "place white at D 4 again" from a target male talker in the GRID corpus [34] (top left) which is mixed with the sentence "lay red in J 9 now" from a masker female talker at -9 dB SIR (top right), which corresponds to the audio file "t2_pwad4a_m4_lrij9n.wav" from the SSC data [27]. It is evident from the figure that the spectro-temporal information of the male talker is masked by the female talker to a great extent. Even without considering the spectrogram for the clean signal (top left), the harmonic structure seen from the spectrogram of the mixture, especially in the first half, suggests that the mixture is dominated by a female talker's voice. The estimated T-F soft mask obtained from the proposed speaker verification based approach is shown in the bottom left panel of the figure along with the spectrogram of the reconstructed target signal (bottom right). Clearly, the interfering masker is removed from the mixture and the reconstructed spectrogram resembles that of the original unmixed spectrogram, albeit at the expense of introducing distortion in the time-frequency content of the target talker's speech. This is not surprising given the extremely low operating SIR for this example. Informal listening experiments suggest good quality of the reconstructed signals, in particular when the masker is of opposite gender.

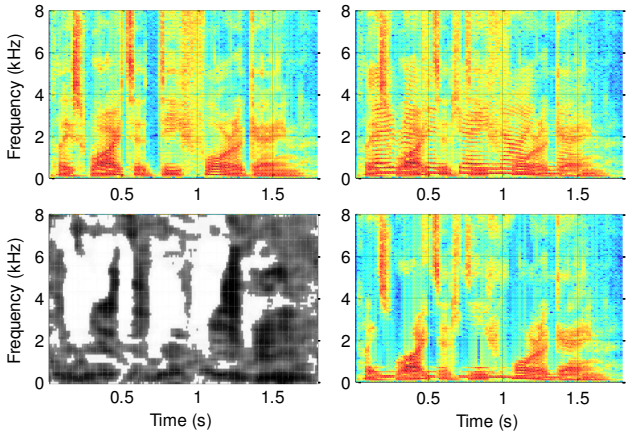

Fig. 2. Sample spectrograms for the sentence "place white at D 4 again" from a target male talker in the GRID corpus [34] (top left) which is mixed with the sentence "lay red in J 9 now" from a masker female talker at -9 dB SIR (top right). The estimated T-F soft mask is shown in the bottom left panel along with the spectrogram of the reconstructed target signal (bottom right).

3. EXPERIMENTS

The proposed speaker verification based masking techniques for co-channel speech recognition are evaluated on the SSC database [27] comprising audio recordings from a total of 34 speakers (16 female and 18 male talkers). Speech material in the SSC database are extracted from the GRID corpus [34] which consists of sentences that follow a fixed and simple grammatical structure L = ⟨command:4⟩ ⟨color:4⟩ ⟨preposition:4⟩ ⟨letter:25⟩ ⟨number:10⟩ ⟨adverb:4⟩, where the numbers in brackets indicate the number of choices at each point. The letter "w" is not included since it is the only multi-syllabic spoken letter in English.

For training, there are 500 clean recordings available per talker, and for tests 600 sentence pairs are mixed at seven different signal-to-interference ratio (SIR) levels: 6, 3 ,0, -3, -6, -9 dB, and clean. The clean test set contains original unmixed recordings from target talkers, virtually representing a condition with infinity (inf) SIR. Within each test set there are approximately equal number of sentence pairs from talkers of opposite genders (DG), talkers of the same gender (SG), and the same talkers (ST). In each sentence pair, the color keyword of the target talker is "white" while that of the masker is not, and the task is to report keyword recognition performance of the letter and number in the target utterance.

In our experiments, HTK is used to perform ASR. All recordings are parameterized into 39-dimensional PNCCs that have been previously shown to be robust alternatives to MFCCs for co-channel speech recognition [29]. For SI acoustic modeling, data from all 34 talkers (i.e., 17000 utterances) are pooled to train context-dependent tied state triphone HMMs. The probability distribution in each state is modeled via a 32-component GMM with diagonal covariance matrices. Maximum *a posteriori* (MAP) adaptation is employed to estimate speaker-dependent (SD) acoustic models using 500 training utterances for each talker. A bi-gram language model reflecting the grammatical structure defined in L is built and used along with the acoustic models to find the most likely state sequence given the observations.

In order to obtain speaker verification likelihoods for the time-level soft masking method (SM-T), a 256-component SI GMM (a.k.a.

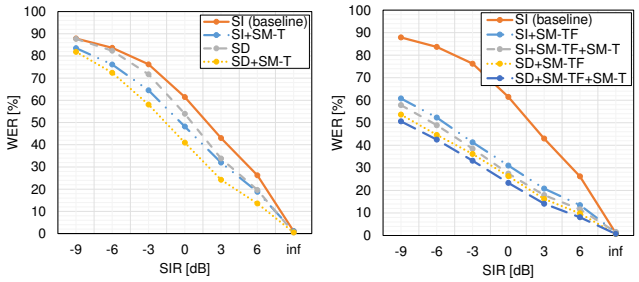

Fig. 3. Co-channel speech recognition performance on the SSC task for the baseline SI and SD systems as well as for systems with SM-T and SM-TF processing methods individually and in combination.

UBM) is learned using training data from all talkers in SSC corpus. Speaker-specific models are then MAP adapted from the UBM. Here, we also use PNCCs as acoustic features. The sigmoid function parameters, b and θ, are set to 1 and -0.25, respectively (for a sensitivity analysis of performance with respect to these parameters see [5]). As for the time-frequency level soft masking approach (SM-TF), a 1024-component UBM is trained on AMS features extracted from T-F units for all talkers, and SD models are estimated using MAP adaptation. We have found that adapting all GMM hyperparameters, i.e., mixture priors, mean vectors, and covariance matrices, is necessary to achieve the best speech separation performance. For the SM-TF method, the sigmoid parameters, b and θ, are set to 1 and 0, respectively. Gaussian models in both methods use diagonal covariance matrices. The MSR Identity Toolbox [35] is used for all our speaker verification experiments.

4. RESULTS

Fig. 3 shows co-channel speech recognition performance on the SSC task for the baseline SI and SD ASR systems as well as for systems with speaker verification based frame-level and time-frequency level soft masking techniques denoted as SM-T and SM-TF, respectively. Results are reported in terms of percent word-error rate (WER), taking into calculations all words in the vocabulary and grammar (for keyword recognition performance see Fig. 4). Several observations can be made from this figure. First, recognition performance of all systems degrades rapidly as the SIR decreases, and performance drop is the largest for the baseline SI system. Second, SD modeling has a small positive impact on the performance, particularly at lower SIRs. This is not surprising given that the SD system has limited ability to distinguish between target-dominant and masker-dominant regions. Third, when the frame-level speaker verification scores are used as reliability measures in the Viterbi search (i.e., the SM-T method), the performance improves significantly. Further improvements in performance are achieved with SD modeling and the SM-T processing. Fourth, the SM-TF processing results in much larger gains in speech recognition performance of co-channel speech. This is expected because in the SM-TF method the processing resolution is much finer compared to the SM-T approach, the target-dominant regions are identified not only at specific time frames, but also certain frequency bands. Finally, combination of the two soft masking techniques yields the greatest boosts in the performance, with relative improvements (reduction in WER) as large as 43% percentage points.

Fig. 4 presents performance comparison of the combined approach (SM-T+SM-TF) versus other strategies that have been evaluated on the SSC task using HTK [17, 18, 19, 20, 22, 23, 36, 37, 38, 39]. Results are given in terms of average keyword (i.e., letter and digit) recognition accuracy across DG and SG test conditions at different SIRs. The ST condition is not considered here because our solutions are based on speaker verification and therefore not applicable to this specific condition. Note, however, that the ST condition (i.e., speakers overlapping with themselves) is unlikely to occur in practice, especially in mobile voice-enabled applications. It is clear that the proposed combined technique compares favorably to other techniques proposed for co-channel speech recognition on the SSC data for SIRs greater than -3 dB. Also, it is seen that our method introduces the least processing artifacts providing the best recognition score on clean test set. This is important because the performance should generalize to other conditions as well.

5. CONCLUSION

This paper has presented two different soft masking strategies based on speaker verification to improve performance of co-channel speech recognition, assuming a known target talker scenario. The speaker verification soft masks were estimated at both time and time-frequency levels and applied to suppress the competing talker interference. It was shown that the proposed methods improve the performance of a baseline "do nothing" ASR system on the SSC data, both individually and in combination. In addition, the combined solution was shown to outperform other techniques evaluated on the same data using HTK. The performance can be further improved using more effective T-F unit representations as well as more powerful classifiers such deep neural networks (DNN).

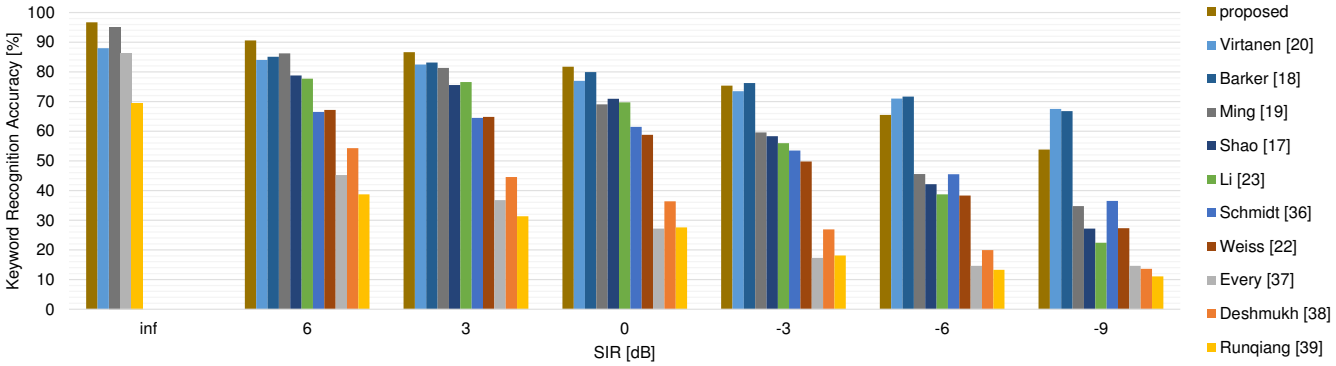

Fig. 4. Performance comparison of different solutions for co-channel speech recognition on the SSC task. Bars show average percent keyword recognition accuracy across DG and SG test conditions at different SIR levels. All compared solutions use HTK in their ASR experiments.

6. REFERENCES

[1] J. A. Naylor and S. F. Boll, "Techniques for suppression of an interfering talker in co-channel speech," in *Proc. IEEE ICASSP*, Dallas, TX, Apr. 1987, pp. 205–208.

[2] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 1, pp. 56–69, Jan. 1990.

[3] D. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 407–424, Sept. 1997.

[4] B. Y. Smolenski, R. E. Yantorno, D. S. Benincasa, and S. Wenndt, "Co-channel speaker segment separation," in *Proc. IEEE ICASSP*, Orlando, FL, May 2002, pp. I–125–I–128.

[5] L. P. Heck and M. Z. Mao, "Automatic speech recognition of co-channel speech: Integrated speaker and speech recognition approach," in *Proc. INTERSPEECH*, Jeju Island, Korea, Oct. 2004, pp. 829–832.

[6] Y. Shao and D. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 289–298, Jan. 2006.

[7] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–1109, 2001.

[8] P. Assmann and A. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Eds. New York: Springer-Verlag, 2004.

[9] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, pp. 233–236, May 2012.

[10] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, Jan. 2010.

[11] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. Elect. Eng., Stanford University, Aug. 1985.

[12] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sounds*. Cambridge, MA: MIT Press, 1990.

[13] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, Oct. 1994.

[14] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. Elect. Eng., Massachusetts Institue of Technology, Jun. 1996.

[15] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[16] S. T. Roweis, "One microphone source separation," in *NIPS 13*. MIT Press, 2000, pp. 793–799.

[17] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 77–93, Jan. 2010.

[18] J. Barker, N. Ma, A. Coy, and M. P. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 94–111, Jan. 2010.

[19] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/independent modeling for speech separation," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 67–76, Jan. 2010.

[20] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 1–4.

[21] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical model approach," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[22] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 16–29, Jan. 2010.

[23] P. Li, Y. Guan, S. wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[24] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*, 2nd ed. Boca Raton, FL: CRC Press, 2013.

[25] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267 – 285, Jun. 2001.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[27] M. P. Cooke and T.-W. Lee. Speech separation challenge website. [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm

[28] S. Young *et al.* HTK - Hidden Markov Model Toolkit v3.4.1. [Online]. Available: http://htk.eng.cam.ac.uk/

[29] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4101–4104.

[30] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 3189–3192.

[31] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Proc. Workshop on Mathematical Foundations of Natural Language Modeling*. Springer, 2002, pp. 105–114.

[32] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1593–1602, 1994.

[33] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, Sept. 2009.

[34] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[35] S. O. Sadjadi, M. Slaney, and L. P. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," in *IEEE SLTC Newsletter*, Novembers 2013. [Online]. Available: http://research.microsoft.com/jump/203749

[36] M. N. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 101–104.

[37] M. R. Every and P. J. B. Jackson, "Enhancement of harmonic content of speech based on a dynamic programming pitch tracking algorithm," in *Proc. INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 101–104.

[38] O. Deshmukh and C. Espy-Wilson, "Modified phase opponency based solution to the speech separation challenge," in *Proc. INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 101–104.

[39] H. Runqiang, Z. Pei, G. Qin, Z. Zhiping, W. Hao, and W. Xihong, "CASA based speech separation for robust speech recognition," in *Proc. INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 101–104.