Directly comparing the listening strategies of humans and machines

Michael I Mandel[1]

[1]Brooklyn College, CUNY, Department of Computer and Information Science

mim@sci.brooklyn.cuny.edu

Abstract

In a given noisy environment, human listeners can more accurately identify spoken words than automatic speech recognizers. It is not clear, however, what information the humans are able to utilize in doing so that the machines are not. This paper uses a recently introduced technique to directly characterize the information used by humans and machines on the same task. The task was a forced choice between eight sentences spoken by a single talker from the small-vocabulary GRID corpus that were selected to be maximally confusable with one another. These sentences were mixed with "bubble" noise, which is designed to reveal randomly selected time-frequency glimpses of the sentence. Responses to these noisy mixtures allowed the identification of time-frequency regions that were important for each listener to recognize each sentence, i.e., regions that were frequently audible when a sentence was correctly identified and inaudible when it was not. In comparing these regions across human and machine listeners, we found that dips in noise allowed the humans to recognize words based on informative speech cues. In contrast, the baseline CHiME-2-GRID recognizer correctly identified sentences only when the time-frequency profile of the noisy mixture matched that of the underlying speech.

Index Terms: Noise, Speech perception, Sentence recognition, Automatic speech recognition

1. Introduction

Despite recent advances in automatic speech recognition (ASR), state-of-the-art systems still suffer from error rates in the range of 30–40% in noisy conditions like personal video recordings [1]. Normal-hearing human listeners are remarkably good at understanding speech in noise, much better than ASR systems [2, 3, 4], even without any grammatical or linguistic information at all [5, 6]. The reasons for these differences, however, are not well understood, and understanding them would very likely directly lead to improvements in ASR noise robustness. Towards this goal, several projects have endeavored to improve ASR noise robustness by building confidence measures of recognition hypotheses based on understanding the errors the recognizer makes and its state when making them [7, 8]. Others have created synthetic data according to various statistical assumptions made in ASR systems [9, 10, 11], estimating the proportion of errors caused by each assumption. And others [12] have applied neurophysiological techniques to a deep neural network acoustic model to try to understand its similarities to human speech perception in quiet environments.

We have recently introduced a method that can reveal the strategy that a listener uses in recognizing a particular utterance in noise [13, 14]. By strategy, we mean the combination of time-frequency "glimpses" [2] that a listener utilizes to recognize a particular utterance when mixed with a particular noise instance in the context of a particular task. While our previous work has used this technique to analyze human speech perception, it can be used with any listener. This paper is the first to use it to directly compare the cues used by humans and ASRs in recognizing specific utterances in noise, and does so in the context of the second CHiME challenge, track 1 (GRID corpus) task [15]. It is also the first to apply this technique to running sentences instead of isolated words, which is not possible with similar approaches, e.g., [16].

Table 1: Sentences selected from the GRID corpus for use in the listening test. All sentences were spoken by talker 16, a female.

ID	Verb	Color	Prep	Let	Num	Adv
BBIKZA	"Bin	blue	in	K	zero	again."
BGIL8A	"Bin	green	in	L	eight	again."
BRIE2A	"Bin	red	in	E	two	again."
BRIK6A	"Bin	red	in	K	six	again."
BRIRZA	"Bin	red	in	R	zero	again."
BWIE8A	"Bin	white	in	E	eight	again."
BWIL2A	"Bin	white	in	L	two	again."
BWIR6A	"Bin	white	in	R	six	again."

The core idea of our technique is to measure the intelligibility of a single recording of an utterance mixed with many different instances of time- and frequency-varying noise. Mixtures in which the utterance is intelligible must have revealed a sufficient amount of information from that utterance for the listener to correctly distinguish it from alternatives. Mixtures in which it is not intelligible, must not have revealed sufficient information. Thus time-frequency regions that are frequently audible in intelligible mixtures and inaudible in unintelligible mixtures are likely to represent the location of important cues that the listener is using. By measuring the correlation between audibility of each time-frequency point with the overall intelligibility of the utterance across mixtures, we can compute the importance of each time-frequency point, which we call the time-frequency importance function (TFIF).

2. Task

The utterances for this experiment were selected from the GRID corpus [17]. The corpus consists of six-word sentences of the form: ⟨*verb*⟩ ⟨*color*⟩ ⟨*preposition*⟩ ⟨*letter*⟩ ⟨*number*⟩ ⟨*adverb*⟩, such as, "Bin blue in K zero again." Each position in the sentence has a fixed number of possible entries: 25 letters (excluding W), 10 digits (including "Zero"), and four words in each of the other positions. Each of 34 talkers recorded 1000 sentences, covering all combinations of colors, letters, and digits, and half of the combinations of the other three words. These talkers represent a wide variety of regional British accents.

We selected this corpus because it facilitates both human and ASR experiments. For human experiments, the corpus provides low predictability from one word to another in the sentences. Thus testing the identification of one sentence in noise to a large

Table 2: Confusion matrices for human and ASR responses. Note that they are responding to different tokens, with significantly more noise in the mixtures presented to the humans. Sentences are abbreviated here to three characters: color, letter, number.

	Human response									ASR response								
True ID	BKZ	GL8	RE2	RK6	RRZ	WE8	WL2	WR6	Sum	BKZ	GL8	RE2	RK6	RRZ	WE8	WL2	WR6	Sum
BKZ	256	30	18	26	43	12	6	9	400	235	5	8	101	18	13	7	13	400
GL8	23	273	16	12	6	39	23	8	400	26	302	3	32	12	3	11	11	400
RE2	4	6	229	41	52	19	42	7	400	6	9	197	36	45	84	12	11	400
RK6	6	6	28	269	21	27	10	33	400	6	24	179	148	6	24	9	4	400
RRZ	10	8	33	20	272	13	22	22	400	15	14	9	78	208	0	15	61	400
WE8	4	15	24	12	15	242	39	50	401	4	10	8	5	8	265	45	55	400
WL2	7	7	38	12	20	27	253	35	399	17	22	7	7	3	73	176	95	400
WR6	5	8	7	54	10	23	42	250	399	10	43	8	6	7	19	203	104	400
Sum	315	353	393	446	439	402	437	414	3199	319	429	419	413	307	481	478	354	3200

extent tests the identification of each of the words in it individually in parallel. The words also provide a good balance of phonetic material. One downside of the corpus for our purposes is that the talkers are British and our listeners are American, making the task slightly more difficult and less natural than if it had been recorded by American talkers.

For ASR experiments, GRID provides a large training corpus for building recognizers. This combined with a small vocabulary (50 words total) makes acoustic models easy to train. In addition, there is a baseline recognizer for the challenge distributed with the Kaldi speech recognition toolkit [18], which we utilized. One downside of using the GRID corpus for our ASR experiments is that this baseline system does not include the use of a deep neural network acoustic model, which has recently become standard in noise-robust recognizers [19].

From these 34,000 utterances, we selected eight to use in the listening tests. Our goal in selecting these sentences was that words in each position be as balanced as possible and as independent as possible from words in the other positions. There was no set of sentences from a single talker in GRID that perfectly satisfied these characteristics, so we selected the set that came as close as possible. The set size of eight sentences was chosen so that an individual subject could perform the entire experiment in a single listening session. The correlational analysis described in Section 2.3 requires approximately 200 mixtures of each utterance. These 200 mixtures will take a human 10–15 minutes to listen to and label. Thus an experiment utilizing eight utterances should take 80–120 minutes.

This selection process resulted in the sentences listed in Table 1, spoken by talker 16, a female. As can be seen in the table, all of the sentences share the same verb, preposition, and adverb. There are four letters, each appearing in two sentences, and four number, each also appearing in two sentences, both meeting our goals for balance. There are also four colors, but two of them appear in a single sentence each and two appear in three sentences each, not meeting the balance goal, but allowing the words in the other positions to do so.

Each sentence was mixed with many instances of "bubble" noise [13]. This noise was designed to provide glimpses of the speech only in specific time-frequency areas, which we call bubbles. To construct this noise, we began with speech-shaped noise with a signal-to-noise ratio (SNR) of -24 dB, sufficient to make the speech completely unintelligible. The noise was then attenuated in "elliptical" bubbles (more accurately described as jointly parabolic in time and ERB-scale frequency [20]), providing glimpses of the speech in these regions. Within each bubble, the noise was suppressed by up to 80 dB. The bubbles were 350 ms wide at their widest and 7 ERB high at their highest, the smallest values that would avoid introducing audible artifacts. The center points of the bubbles were selected uniformly at random in time and in ERB-scale frequency, except that they were excluded from a 2-ERB buffer at the bottom and top of the frequency scale to avoid edge effects.

2.1. Human listening test

In the human listening test, one sentence was selected at random, mixed with bubble noise, and presented to the listener, who then selected one of the eight possibilities. Sentence presentation was blocked, so that every block of eight mixtures used each of the sentences once in a random order. The number of bubbles per sentence controls the difficulty of the task, and was adapted using the weighted up-down procedure [21] separately for each sentence starting at 30 bubbles per sentence. When a sentence was correctly identified, the number of bubbles used in its next presentation was reduced by 2% and when it was incorrectly identified, the number of bubbles used in its next presentation was increased by 2.3%. This asymmetry leads the procedure to converge to the number of bubbles per sentence that allows the listener to correctly identify 56.3% of the mixtures, half way between chance and perfect performance. This procedure resulted in a final bubble rate of 18–24 bubbles per sentence, varying by listener and utterance.

The human listening test was performed over headphones via a MATLAB interface. Subjects consisted of one expert listener, who labeled 1600 mixtures and was familiar with bubble noise, and three naïve listeners, who together labeled another 1600 mixtures (401, 562, and 639 mixtures each) and had never heard bubble noise before. All were native speakers of American English. Subjects were allowed to familiarize themselves with the clean utterances and the task for 5 minutes before noise was added. They were allowed to adjust presentation volume to a comfortable level, listen to each mixture as many times as they wanted, take breaks regularly, and end their participation whenever they wanted. Feedback was provided only at the end of the training period and no feedback was provided during the experiment itself. The listeners each spent approximately the same amount of time on the test, the expert being the fastest.

2.2. ASR listening test

The ASR system used in these experiments was based on the Kaldi baseline for the first CHiME challenge [22] and also the second CHiME challenge, track 1 [15]. The training data consisted of speech from the GRID corpus mixed with various noises recorded in a household environment. The recognizer used a Gaussian mixture model (GMM) front end operating on mel-frequency cepstral coefficients (MFCCs) predicting clustered triphone states. The MFCCs were transformed using linear discriminant analysis of several consecutive frames followed by a global maximum likelihood linear transformation [23]. It used

2500 triphone states and 15000 Gaussians. These GMMs were trained on the training data from the second CHiME challenge, track 1, which consists of 17,000 noisy utterances, 500 from each talker. Our training excluded utterances from talker 16, the one used in the listening test.

Several modifications to the decoding parameters were necessary to perform the same listening task as the human listeners. First, we modified the grammar to consist of only the eight test sentences as eight parallel paths from the start state to the end state. After doing so, we needed to modify the weights on each of the sentence paths in the grammar to achieve approximate parity in the frequency with which each sentence was selected. This required placing a large penalty (52 nats, where a nat is a unit of information like a bit, but using the natural logarithm) on selecting BWIE8A, moderate penalties (46, 43, 39, and 36 nats) on selecting BGIL8A, BRIE2A, BRIRZA, and BWIL2A, respectively, and low penalties (16 and 0 nats) on selecting BWIR6A and BRIK6A. Apparently in bubble noise the recognizer was particularly unlikely to select the utterances containing the word "Six". We also found that with the default settings, many sentences' transcripts ended before the final state due to beam search starvation. Because this is a small-vocabulary task, we were able to increase the width of the beam to 200 nats to eliminate this issue, presumably by exhaustively exploring all paths.

As expected, in order to correctly identify 50% of the sentences, the ASR could only tolerate much milder bubble noise than the human listeners. We performed several searches across the number of bubbles per sentence and SNR to identify a good operating point, settling on -9 dB SNR before adding bubbles and 54 bubbles per sentence. This resulted in an actual average SNR of -2.9 dB. One advantage of the ASR listener over the human listeners is that it can listen faster than real time and has an unlimited attention span. We thus utilized 400 mixtures per utterance with it.

2.3. Analysis technique

In order to analyze the results, we compute the point-biserial correlation between the dichotomous variable y_{ij}, whether or not the listener correctly identified the jth mixture of the ith utterance, and the continuous variable $N_{ij}(\omega, t)$, the audibility of time-frequency point (ω, t) in the jth mixture of the ith utterance. Audibility here is defined as the proportion of attenuation (in dB) applied to the noise at that point, i.e., the depth of the bubble, ranging between 0 for no attenuation (pure noise) and 1 for total attenuation (no noise). This correlation is performed across mixtures, but separately for each time-frequency point for each utterance, leading to a "massively univariate" correlation, denoted $c_i(\omega, t)$. The significance of this correlation can be assessed using a one-way ANOVA with two levels, resulting in a p-value of $p_i(\omega, t)$ for each point and utterance. The visualizations in Figure 1 shows the the spectrogram of each utterance with lightness (in the HSV color space) set to $0.5 + 0.5 \exp(-p_i(\omega, t)/0.05)$, so that significant correlations are shown at full lightness and insignificant correlations are shown at half lightness.

While the task itself is an eight-way forced choice between sentences, these choices can also be analyzed at the word level. For example, if a sentence is mistaken for one the shares the same letter word, then this can be considered a correct selection for analyzing the importance of cues to that letter word. Thus the same responses to the same stimuli can be interpreted as having several different meanings for the purposes of our analyses, similarly to the information transmission analysis of [24]. TFIFs based on these different interpretations are shown as different rows in Figure 1.

3. Human results and discussion

Time-frequency importance functions derived from the human responses are shown in the top half of Figure 1. It shows that the human listeners are attending to time-frequency locations in the spectrogram corresponding to various speech cues. These cues include the initial glides of "white" and "red", the initial stop burst of "two", and the initial sibilance of "six". These results are consistent across different productions of the same word. They also follow the well-established cues of speech production for these words [25] and agree with other analyses of cues for speech perception of individual tokens [16]. The identified regions, however, only include a subset of the distinctive features that might be expected. For example, the final sibilance in "six" does not appear to be utilized consistently, even though it is nearly as loud as the initial sibilance in BWIR6A. Nor is the initial sibilance in "zero" utilized, although low frequency information in the /z/ does appear to be utilized.

Another interesting feature of these results is that different "correctness" signals (different rows in Figure 1) show correlations with different time-frequency regions of the utterances. This analysis is possible because of the use of sentence stimuli, in contrast to previous auditory bubbles experiments, which employed isolated words [13, 14]. For word-level correctness, these correlations generally appear in the spectral regions of the word in question, an effect that is very noticeable in all of the "red" and "white" sentences, but especially for the sentences BRIE2A, BRIRZA, and BWIL2A. For example, in BRIRZA, when the correctness of identifying the color word "red" is considered, the importance is high in the region of the second formant transition of the /r/ in "red", which occurs around 300 ms into the utterance. When correctness for the letter word "R" is considered instead, the importance shifts to the second formant transition into and during that word, around 600 ms. And when correctness for the number word "Zero" is considered, the importance shifts to the first formant of the initial /z/, around 900 ms.

Frequently, however, the importance for one word includes regions of other words. For example, in BWIR6A, the importance regions for correctly recognizing the word "R" include both the formant transition of the /w/ in "White" and the sibilance of the /s/ in "Six". This could be explained by the fact that correctly identifying the words "White" and "Six" uniquely identifies this sentence, regardless of whether "R" was audible. Similarly, for the two sentences with unique colors (BBIKZA and BGIL8A), the importance tends to focus on just the color words, even for correct identification of other word positions.

The TFIFs for sentence-level correctness generally correspond to the union of the TFIFs for the word-level correctness for a given utterance. This is most noticeable for the sentences BRIRZA, BRIE2A, and BWIR6A. Often, however, this sentence-level correctness shows a weaker correlation than the word-level. This is likely because many cues can be used to identify a sentence, so no single cue is audible in all correct identifications.

4. ASR results and discussion

Time-frequency importance functions derived from the ASR responses are shown in the bottom half of Figure 1. It shows that the ASR is mainly "attending" to regions where there is little speech energy. For example, there is a large importance region in the low energy high frequencies above the end of the word

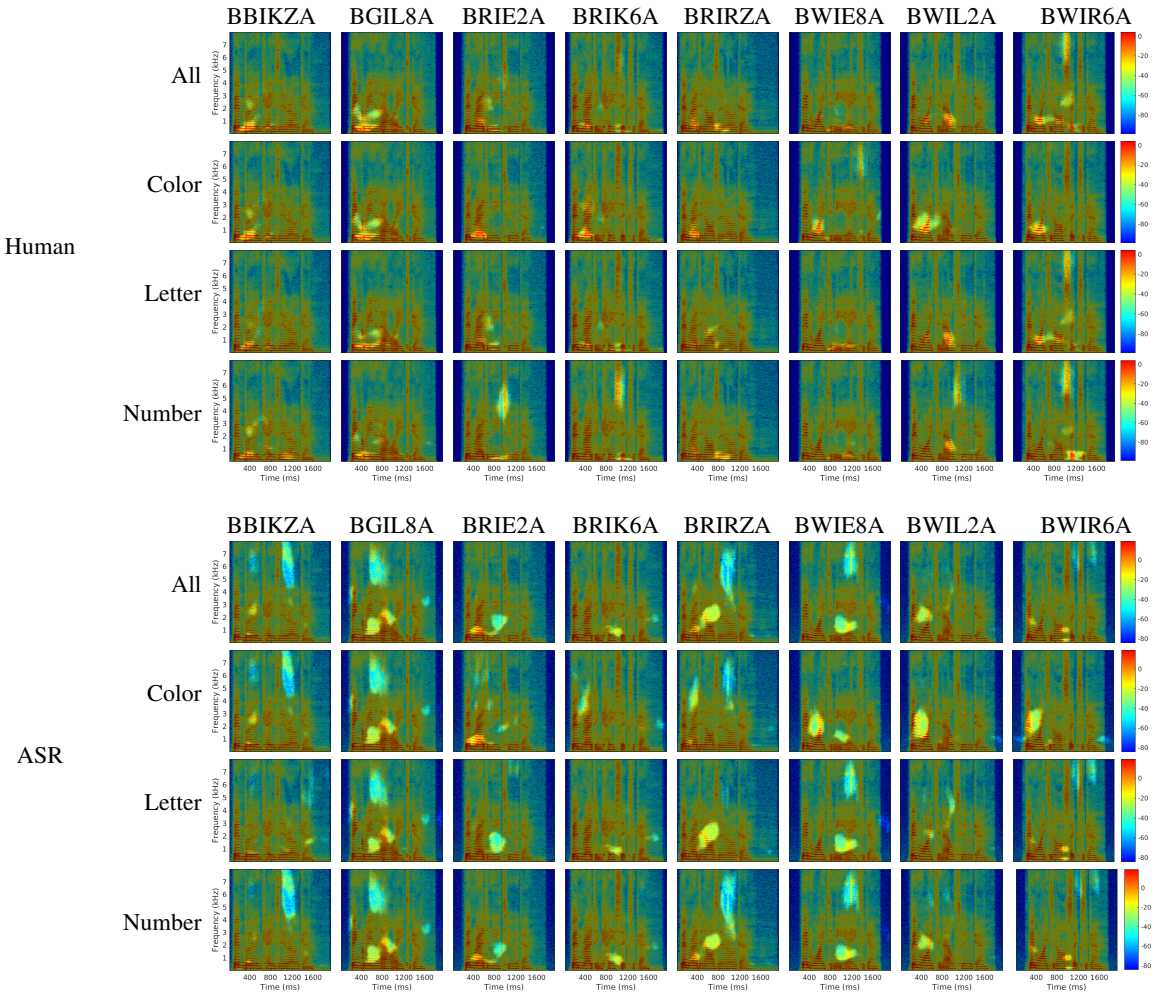

Figure 1: Time-frequency importance functions for human and automatic speech recognition on all eight sentences using four different "correctness" signals (one per row): Sentence-level correctness ("All"), Color word, Letter, and Number.

"Zero" in BRIRZA. Similarly, there is importance in the high frequencies above "E Eight" in BWIE8A and above "Green In" in BGIL8A. Interestingly, there is also importance between the first two formants in "E" in BRIE2A, a low-frequency region of low energy. The ASR TFIFs appear much less sensitive to the type of correctness under consideration (i.e., the ASR rows of the figure are quite similar).

The ASR importance does include some formant transitions, although it seems to be focusing more on the lack of energy adjacent to the formants as opposed to the high energy of the formants themselves. For example, there are large importance regions above the second formant transition in "In R" in BRIRZA, after the falling second formant of "L" in BGIL8A, and before the rising second formant of "White" in all of the sentences that include it. One counterexample is the meeting of the first and second formants in the /r/ of "Red" in BRIE2A, which is both energetic and important to the ASR directly.

Overall these results suggest that the ASR can correctly identify a word or sentence when the noise that is added to it has a similar spectral profile to the speech itself. The ASR is thus using gaps in the noise very differently from the human listeners. While the humans use gaps to identify speech that is revealed, the ASR uses the general spectral shape of the mixture to identify the speech. This is very likely a result of the ASR using a GMM-based acoustic model with MFCC features. The MFCCs characterize the gross spectral shapes of the entire mixture, and cannot separate the speech from the noise at all. In the future, we will perform this same analysis on recognizers that are better able to separate speech from noise, including those using deep neural network (DNN) acoustic models [19] operating on log mel spectra, explicit noise suppression or source separation [26], and missing-data approaches [27].

5. Conclusions

This paper has described an experiment to directly compare human listeners with automatic speech recognizers in terms of their strategies for recognizing speech in noise. It has shown that in this task, humans focus on time-frequency regions corresponding to formant transitions, stop bursts, and sibilance, but a traditional GMM-based acoustic model operating on MFCCs seems to focus on regions of low energy. Thus for the humans, low energy regions of noise revealed speech and allowed them to correctly identify it, while for the ASR system, these glimpses were not useful. Instead, it could only correctly identify mixtures with a similar spectral profile to the clean speech. The paper has also shown that the auditory bubbles technique [13, 14] can operate just as well on running sentences as isolated words and on ASR listeners just as well as human listeners.

6. References

[1] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Proc. ASRU*. IEEE, Dec. 2013, pp. 368–373.

[2] M. P. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, pp. 1562–1573, 2006.

[3] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, Jul. 1997.

[4] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, May 2007.

[5] B. Meyer, T. Wesker, T. Brand, A. Mertins, and B. Kollmeier, "A human-machine comparison in speech recognition based on a logatome corpus," in *Proc. Workshop on Speech Recognition and Intrinsic Variation*, 2006.

[6] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. EL256–EL261, Mar. 2012.

[7] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools," in *Proc. IEEE ICASSP*, vol. 1, 1995, pp. 221–224.

[8] L. Chase, "Error-responsive feedback mechanisms for speech recognizers," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1997.

[9] D. Gillick, L. Gillick, and S. Wegmann, "Don't multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition," in *Proc. ASRU*. IEEE, Dec. 2011, pp. 71–76.

[10] D. Gillick, S. Wegmann, and L. Gillick, "Discriminative training for speech recognition is compensating for statistical dependence in the HMM framework," in *Proc. IEEE ICASSP*. IEEE, Mar. 2012, pp. 4745–4748.

[11] S. H. K. Parthasarathi, S.-Y. Chang, J. Cohen, N. Morgan, and S. Wegmann, "The blame game in meeting room ASR: An analysis of feature versus model errors in noisy and mismatched conditions," in *Proc. IEEE ICASSP*. IEEE, May 2013, pp. 6758–6762.

[12] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Proc. Interspeech*, 2015.

[13] M. I. Mandel, "Learning an intelligibility map of individual utterances," in *Proc. IEEE WASPAA*, 2013.

[14] M. I. Mandel, S. E. Yoho, and E. W. Healy, "Generalizing time-frequency importance functions across noises, talkers, and phonemes," in *Proc. Interspeech*, 2014.

[15] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE ICASSP*, May 2013, pp. 126–130.

[16] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, Apr. 2010.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[19] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE ICASSP*, 2013.

[20] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.

[21] C. Kaernbach, "Simple adaptive testing with the weighted up-down method," *Perception & Psychophysics*, vol. 49, no. 3, pp. 227–229, may 1991.

[22] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, may 2013.

[23] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Tr. SAP*, vol. 7, no. 3, pp. 272–281, may 1999.

[24] G. Miller and P. Nicely, "An analysis of perceptual confusions among some english consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, Mar. 1955.

[25] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics. MIT Press, 2000.

[26] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba *et al.*, "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," in *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011, pp. 12–17.

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