Comparative Evaluation of Speech Enhancement Methods for Robust Automatic Speech Recognition

Kuldip K. Paliwal, James G. Lyons, Stephen So, Anthony P. Stark, and Kamil K. Wójcicki

Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane Queensland 4111, Australia
{k.paliwal, j.lyons, s.so, a.stark, k.wojcicki}@griffith.edu.au

Abstract

A comparative evaluation of speech enhancement algorithms for robust automatic speech recognition is presented. The evaluation is performed on a core test set of the TIMIT speech corpus. Mean objective speech quality scores as well as ASR correctness scores under two noise conditions are given.

Index Terms: Speech enhancement, robust ASR

1. Introduction

When trained and tested under similar (matched) conditions, the current state-of-the-art automatic speech recognition (ASR) systems perform reasonably well. Their performance, however, drops significantly under mismatched conditions, i.e. when training is performed on clean speech, while testing is done on noisy speech. Various approaches have been discussed in the literature for robust ASR under mismatched conditions. One approach is to use speech enhancement as a pre-processor to ASR, where a speech enhancement algorithm is applied on corrupted speech prior to feature extraction. A conventional ASR trained on clean speech is then used in conjunction with features extracted from the enhanced speech. A second approach is to use robust feature extraction methods which utilize perceptual properties of the human ear to increase performance in the presence of noise and distortion. Typical examples of robust feature extraction algorithms include Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) coefficients [2]. A third approach to robust ASR is to process the noisy features in some way prior to recognition. Typical examples of feature enhancement include RASTA filtering [3] and cepstral mean subtraction (CMS) [4]. A fourth approach is through model adaptation, where a clean model can be adapted to match the noisy conditions. Alternatively, the noisy features can be adapted to fit the original clean model. Parallel Model Combination (PMC) [5] is an example of a model adaptation approach. Some combinations of the above approaches have also been proposed [6, 7].

In the present paper our aim is to investigate the role of speech enhancement as a pre-processor for robust ASR. The aim of speech enhancement is to improve the quality of noisy speech so that it is suitable for human listeners. At the same time, these algorithms should at least preserve speech intelligibility, however, that is rarely the case [8]. Consequently, such methods may do an excellent job as far the quality for the human ear is concerned, but they may not be good for machine recognition. Thus, a straightforward use of speech enhancement methods does not guarantee an improvement over ASR performance under noisy conditions. Our aim in this paper is to identify which particular methods fit well with ASR. In the past various authors have investigated this problem, for example [9, 10, 11, 12, 13, 14]. However, such studies are rarely done on a common database and not all the methods of speech enhancement are included. This present work aims to explore in a unified manner a broad set of speech enhancement techniques to determine their performance for robust ASR.

2. Experiments

Our aim in this paper is to investigate the role of speech enhancement as a pre-processor for robust ASR. In particular, we are interested in which speech enhancement methods are most suited for improving robustness of ASR. For this purpose speech enhancement and ASR experiments are conducted. The reminder of this section describes the these experiments.

2.1. Speech enhancement

In this comparative evaluation we consider 16 commonly employed speech enhancement methods belonging to four major classes of speech enhancement algorithms. Table 1 lists these methods along with references to the original work. All of the algorithms are also described in [15] along with reference implementations. In our experiments we employ these implementations with their default settings. Note that adaptive noise estimation methods are outside of scope of this evaluation. Instead initial five, 20 ms, non-overlapped frames are used for noise estimation. Some methods also use a simple VAD (as per [15]). In addition to ASR evaluation of the above algorithms (described in Section 2.2), we also perform a corresponding objective evaluation of speech quality in terms of mean PESQ scores [16]. The mean PESQ scores are computed across the core test set of the TIMIT corpus [17]. Two additive noise types, white and babble, are investigated. The noise signals are taken from the NOISEX-92 noise database [18].

2.2. Automatic speech recognition

The automatic speech recognition (ASR) experiments were conducted on the TIMIT speech corpus [17]. The TIMIT speech corpus is sampled at 16 kHz and consists of 6300 utterances spoken by 630 speakers. The corpus is separated into training and testing sets. For our experiments the *sa** utterances, which are the same across all speakers, were removed from both the training and testing sets to prevent biasing the results. The ASR training is performed on clean speech, while for testing, clean speech is first corrupted by additive noise and then processed using speech enhancement techniques listed in Table 1. For training we use full train set consisting of 3696 utterances from 462 speakers. For testing we use the core test set consisting of

978-1-4244-7907-8/10/$26.00

Table 1: *List of speech enhancement algorithms evaluated in the present study as pre-processors for the TIMIT ASR task.*

ALGORITHM CLASS	ALGORITHM	REFERENCE
Spectral subtractive	**SSUB**	[19]
	MBAND	[20]
	RDC	[21]
Wiener-type	**Wiener-as**	[22]
	Wiener-wt	[23]
Statistical model-based	**MMSE**	[24]
	MMSE-SPU	[24]
	logMMSE	[25]
	logMMSE-SPU-1	[26]
	logMMSE-SPU-2	[26]
	logMMSE-SPU-3	[27]
	logMMSE-SPU-4	[28]
	STSA-weuclid	[29]
	STSA-wcosh	[29]
Subspace	**KLT**	[30]
	pKLT	[31]

192 utterances from 24 speakers. A HTK-based triphone recognizer with 3 states per HMM and 8 gaussian mixtures per state is used. The phoneme set consisting of 48 phonemes is reduced to 39 for testing purposes as in [32]. The frame size was set to 25 ms with a step size of 10 ms. MFCC features with energy coefficients, as well as the first and second derivatives (39 coefficients total) were used. Cepstral mean subtraction was applied. A bigram language model is used. For recognition, the Viterbi algorithm is used with no pruning factor. The Viterbi decoder used a likelyhood scaling factor of 8 and a penalty of 0. Phoneme recognition results are quoted in terms of correctness percentage (Corr (%)) [33].

3. Results and discussion

The results of speech enhancement as well as ASR experiments are shown in Table 3a and 3b, for white and babble noises, respectively. The results suggest no single choice for ASR speech enhancement – with performance varying substantially across both noise types and input SNRs. The best performing algorithms for each category are summarized in Table 2. Overall, the best performing enhancement methods were Wiener-as (across all SNRs), RDC (high SNRs) and STSA-wcosh (low SNRs). In general, most enhancement algorithms performed quite well, producing modest improvements in ASR performance. One area in which all algorithms performed badly was the clean case (SNR=∞), with all methods showing degradation of ASR performance. However, in the case of MMSE and RDC, these performance drops were quite small. Perhaps employing speech enhancement pre-processing on clean speech prior to training could also be investigated. Notably, improvements in objective speech quality (in terms of mean PESQ scores) did not translate to ASR correctness scores improvements. For the white noise case, the KLT method produced best objective speech quality scores, yet its corresponding ASR performance was quite poor.

Table 2: *Best speech enhancement performers for ASR*

NOISE	LOW SNR (0 dB – 5 dB)	MEDIUM SNR (10 dB – 15 dB)	HIGH SNR (20 dB – 30 dB)
White	STSA-wcosh	MMSE-SPU	Wiener-as
Babble	MMSE-SPU	RDC	RDC

4. References

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

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

[3] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974.

[5] M. Gales, "Model-based techniques for noise robust speech recognition," 1996. [Online]. Available: citeseer.ist.psu.edu/gales95modelbased.html

[6] J. Flores and S. Young, "Continuous speech recognition in noise using spectral subtraction and hmm adaptation," *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. i, pp. I/409–I/412 vol.1, 19-22 Apr 1994.

[7] D. Pei and C. Zhigang, "An efficient robust automatic speech recognition system based on the combination of speech enhancement and log-add hmm adaptation," *Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on*, vol. 3, pp. 367–371 vol.3, 2001.

[8] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Acoustical Society of America Journal*, vol. 122, pp. 1777–+, 2007.

[9] J. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 39, no. 4, pp. 795–805, Apr 1991.

[10] M. Shozakai, S. Nakamura, and K. Shikano, "A speech enhancement approach e-cmn/css for speech recognition in car environments," *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pp. 450–457, 14-17 Dec 1997.

[11] G. Lathoud, M. Magimai.-Doss, B. Mesot, and H. Bourlard, "Unsupervised Spectral Subtraction for Noise-Robust ASR," in *Proceedings of the 2005 IEEE ASRU Workshop*, San Juan, Puerto Rico, December 2005, iDIAP RR 05-42.

[12] R. Gemello, F. Mana, and R. De Mori, "Automatic speech recognition with a modified ephraim-malah rule," *Signal Processing Letters, IEEE*, vol. 13, no. 1, pp. 56–59, Jan. 2006.

[13] A. Abolhassani, S.-A. Selouani, and D. O'Shaughnessy, "Speech enhancement using pca and variance of the reconstruction error in distributed speech recognition," *Automatic Speech Recognition and Understanding, 2007. ASRU. IEEE Workshop on*, pp. 19–23, 9-13 Dec. 2007.

[14] D. Dimitriadis, J. C. Segura, L. Garcia, A. Potamianos, P. Maragos, and V. Pitsikalis, "Advanced front-end for robust speech recognition in extremely adverse environments," *Proc. Interspeech- Eurospeech 2007, Antwerp, Belgium*, vol. 93, Aug. 2007.

[15] P. Loizou, *Speech Enhancement: Theory and Practice*. Taylor and Francis, Boca Raton, FL., 2007.

[16] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'01)*, vol. 2, Salt Lake City, Utah, USA, 2001, pp. 749–752.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 27 403–+, Feb. 1993.

[18] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[19] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'79)*, Washington, DC, USA, 1979, pp. 208–211.

Table 3: *TIMIT experimental results: mean PESQ scores and phoneme correctness (%) scores for (a) white and (b) babble noises. A bold score indicates the best performing method for a given SNR.*

(a)

ALGORITHM	MEAN PESQ								CORR (%)							
SNR (dB)	**0**	**5**	**10**	**15**	**20**	**25**	**30**	∞	**0**	**5**	**10**	**15**	**20**	**25**	**30**	∞
Noisy (white)	1.55	1.90	2.26	2.62	2.97	3.31	3.64	**4.50**	14.11	22.16	33.00	45.31	55.95	64.73	70.52	**76.77**
SSUB	1.78	2.29	2.72	3.17	3.58	3.90	**4.12**	4.36	25.41	36.88	46.25	56.61	63.52	68.83	73.02	74.94
MBAND	1.61	2.08	2.58	2.97	3.21	3.38	3.49	3.60	16.26	25.44	37.84	52.06	60.88	66.61	68.06	69.54
RDC	1.61	2.00	2.41	2.82	3.21	3.58	3.93	4.38	15.75	24.83	36.50	49.82	62.10	70.05	74.06	75.92
Wiener-as	2.01	2.42	2.79	3.12	3.43	3.72	3.99	4.39	30.89	43.21	54.52	63.14	**68.19**	**72.19**	**74.86**	75.27
Wiener-wt	1.78	2.23	2.64	3.07	3.45	3.79	4.03	4.32	8.77	16.76	29.60	44.17	56.23	63.59	69.20	74.47
MMSE	2.04	2.41	2.74	3.05	3.33	3.61	3.87	4.42	23.10	32.69	43.61	55.04	63.67	69.57	73.65	76.52
MMSE-SPU	2.14	2.57	2.93	3.25	3.58	3.87	4.09	4.34	32.57	44.65	**55.19**	**63.52**	68.03	70.87	72.20	74.26
logMMSE	2.13	2.54	2.88	3.17	3.45	3.71	3.96	4.38	27.57	40.05	50.74	60.47	66.93	71.25	73.72	75.52
logMMSE-SPU-1	1.96	2.42	2.84	3.21	3.54	3.80	3.99	4.26	33.25	42.86	52.04	57.20	61.76	65.74	67.98	72.87
logMMSE-SPU-2	1.94	2.39	2.81	3.20	3.52	3.78	3.98	4.25	32.13	42.20	51.87	56.77	61.06	64.85	67.85	72.74
logMMSE-SPU-3	2.11	2.53	2.90	3.25	3.55	3.82	4.03	4.31	32.89	42.98	53.36	59.61	64.98	68.82	70.84	74.35
logMMSE-SPU-4	1.65	2.15	2.54	2.88	3.26	3.60	3.90	4.35	29.47	37.12	46.11	53.73	62.36	68.16	71.57	75.26
STSA-weuclid	2.11	2.52	2.88	3.20	3.51	3.78	4.02	4.34	32.66	42.89	54.30	62.13	66.65	69.67	71.38	74.29
STSA-wcosh	2.15	2.53	2.87	3.20	3.51	3.79	4.02	4.32	**36.06**	**46.08**	54.79	61.81	65.52	68.64	70.93	73.17
KLT	**2.17**	**2.60**	**2.97**	**3.34**	**3.66**	**3.92**	4.10	4.30	20.51	31.24	42.76	51.85	60.58	66.39	70.94	73.88
pKLT	1.97	2.28	2.64	3.02	3.40	3.73	3.99	4.28	32.26	40.30	48.65	56.36	62.39	66.52	69.78	75.61

(b)

ALGORITHM	MEAN PESQ								CORR (%)							
SNR (dB)	**0**	**5**	**10**	**15**	**20**	**25**	**30**	∞	**0**	**5**	**10**	**15**	**20**	**25**	**30**	∞
Noisy (babble)	1.75	2.08	2.43	2.77	3.10	3.43	3.74	**4.50**	24.27	33.32	45.29	56.49	65.16	71.16	73.68	**76.77**
SSUB	1.68	2.13	2.56	2.97	3.36	**3.69**	**3.94**	4.36	24.80	33.98	43.99	53.46	61.05	66.62	70.24	74.94
MBAND	**2.00**	**2.36**	**2.69**	2.98	3.21	3.37	3.47	3.60	31.03	42.00	**52.95**	60.39	64.60	66.87	68.39	69.54
RDC	1.74	2.13	2.53	2.91	3.27	3.60	3.89	4.38	27.71	37.99	50.40	**61.89**	**68.58**	**72.64**	**74.13**	75.92
Wiener-as	1.86	2.24	2.61	2.97	3.31	3.63	3.89	4.39	31.72	40.67	50.91	61.18	66.43	71.06	72.60	75.27
Wiener-wt	1.34	1.88	2.38	2.84	3.27	3.64	3.92	4.32	26.42	36.22	47.59	57.17	64.85	68.32	71.28	74.47
MMSE	1.94	2.28	2.64	2.97	3.28	3.58	3.84	4.42	29.12	38.87	50.34	61.34	67.52	71.01	73.65	76.52
MMSE-SPU	1.92	2.29	2.68	**3.04**	**3.38**	3.69	3.94	4.34	**33.33**	**42.73**	52.48	60.49	65.45	68.72	71.56	74.26
logMMSE	1.94	2.31	2.67	3.02	3.33	3.62	3.88	4.38	32.45	41.78	51.88	61.81	67.38	70.56	73.08	75.52
logMMSE-SPU-1	1.77	2.19	2.59	2.99	3.34	3.64	3.88	4.26	29.76	37.75	49.06	56.80	62.60	66.08	69.36	72.87
logMMSE-SPU-2	1.78	2.19	2.60	2.99	3.34	3.64	3.88	4.25	29.73	37.53	48.16	56.76	62.23	66.13	69.40	72.74
logMMSE-SPU-3	1.81	2.23	2.62	3.01	3.35	3.66	3.90	4.31	30.84	39.69	51.38	59.54	64.91	68.44	71.23	74.35
logMMSE-SPU-4	1.56	2.07	2.53	2.93	3.28	3.59	3.86	4.35	29.25	38.74	49.19	59.96	66.74	70.69	73.09	75.26
STSA-weuclid	1.92	2.29	2.67	3.03	3.36	3.65	3.91	4.34	32.85	42.41	52.32	59.95	65.35	68.80	71.73	74.29
STSA-wcosh	1.86	2.22	2.61	2.98	3.32	3.63	3.89	4.32	31.24	39.67	50.61	58.56	63.53	66.90	70.21	73.17
KLT	1.70	2.16	2.58	2.98	3.35	3.68	3.94	4.30	27.51	35.57	46.80	56.07	62.57	67.34	70.69	73.88
pKLT	1.44	1.94	2.42	2.88	3.28	3.63	3.89	4.28	32.57	41.14	51.91	60.40	66.14	69.93	73.31	75.61

[20] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP'02)*, 2002.

[21] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 8, pp. 799–807, Nov 2001.

[22] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, pp. 629–632 vol. 2, 7-10 May 1996.

[23] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, 2004.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[25] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[26] S. N. K. Ing Yann Soon and C. K. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," *Signal Processing*, vol. 75, no. 2, pp. 151–159, Jun 1999.

[27] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, vol. 2, pp. 789–792 vol.2, 15-19 Mar 1999.

[28] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *Signal Processing Letters, IEEE*, vol. 9, no. 4, pp. 113–116, Apr 2002.

[29] P. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

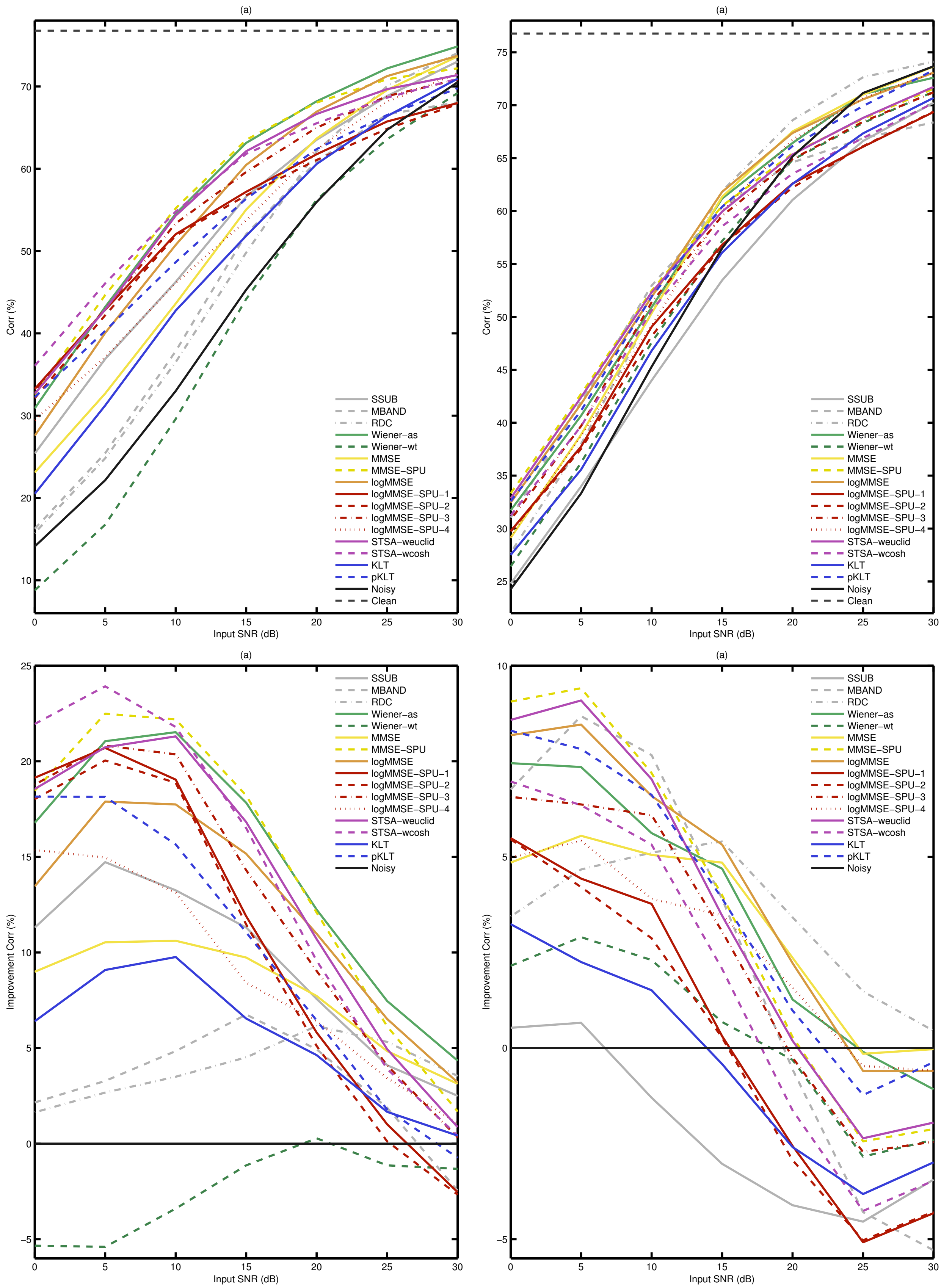

Figure 1: *Phoneme correctness (%) scores versus input SNR (dB) for the TIMIT ASR task.*

[30] Y. Hu and P. Loizou, “A subspace approach for enhancing speech corrupted by colored noise,” *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 1, pp. I–573–I–576 vol.1, 2002.

[31] F. Jabloun and B. Champagne, “Incorporating the human hearing properties in the signal subspace approach for speech enhancement,” *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

[32] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden markov models,” *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[33] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, 3rd ed., Cambridge University Engineering Department, 2006.