



# Adding Noise to Improve Noise Robustness in Speech Recognition

Nicolás Morales<sup>1</sup>, Liang Gu<sup>2</sup> and Yuqing Gao<sup>2</sup>

<sup>1</sup>HCTLab, Universidad Autónoma de Madrid, Spain

<sup>2</sup>IBM T. J. Watson Research Center, Yorktown Heights, USA

nicolas.morales@uam.es, liangu@us.ibm.com, yuqing@us.ibm.com

## Abstract

In this work we explore a technique for increasing recognition accuracy on speech affected by corrupting noise of an undetermined nature, by the addition of a known and well-behaved noise (masking noise). The same type of noise used for masking is added to the training data, thus reducing the gap between training and test conditions, independent of the type of corrupting noise, or whether it is stationary or not. While still in an early development stage, the new approach shows consistent improvements in accuracy and robustness for a variety of conditions, where no use is made of a-priori knowledge of the corrupting noise. The approach is shown to be of particular interest to the case of cross-talk corrupting noise, a complicated situation in speech recognition for which the relative gain with the proposed approach is over 24%.

**Index Terms:** acoustic noise, robustness, speech enhancement, speech recognition

## 1. Introduction

Automatic Speech Recognition (ASR) is known to suffer a significant performance loss when speech is corrupted by noise; a condition that sets a major challenge for many real world applications. In this work we explore a robustness technique consisting in the addition of noise to already noisy speech, with the goal of reducing the mismatch between training and testing conditions.

There are different fields where improvements can be made to improve robustness (these may also be combined): 1) signal parameterization (feature extraction), 2) speech enhancement and 3) model adaptation or retraining. The first one is on the extraction of features robust to noise. Different parameterizations have been studied in the past, among which some of the most successful are MFCC [1], or RASTA [2], and their variations. The other two strategies –speech enhancement and model adaptation or retraining– have a common characteristic in that they aim at reducing the mismatch between acoustic models and input speech caused by a corrupting noise. Enhancement techniques are typically deterministic: they learn from the noisy signal and increase the Signal to Noise Ratio (SNR) based on predictions. This can be done prior to parameterization as in Spectral Subtraction [3], or over the extracted features, as is the case in Cepstral Mean Normalization [4], Wiener filtering, and the more sophisticated SPLICE-like approaches [5]. However, noise is a random process, and thus, deterministic solutions do not seem appropriate. This explains why such techniques are more successful dealing with stationary noises than with non-stationary noises where predictions are unfeasible.

The third family of solutions is acoustic model adaptation to the new environment (typically MLLR [6] or MAP [7]), or model retraining. An extension of the later is multi-style training where different noisy conditions are combined during training for a more generalist modeling. A recently developed technique, uncertainty decoding, combines feature enhancement with real-time acoustic model’s covariance matrix adaptation to exploit the degree of certitude in the enhancement step [8]. Nevertheless, these techniques require a-priori knowledge of the noise source (typically noisy data is required), which is not always available, and are difficult to generalize.

We propose a non-deterministic approach that aims at improving ASR accuracy and providing a general framework for recognition in noisy environments. The idea is depicted in Fig. 1. The original clean speech is corrupted by environmental or channel noise (*corrupting noise*), and in order to reduce the mismatch between training and test conditions created by this noise, another source of well-behaved noise is added to both test data (*masking noise*) and training data (*training noise*). The hope is that the noise artificially added to the input speech will mask the corrupting noise. Adding noise may seem a peculiar way of increasing robustness, but it can be thought of as bringing an unknown condition to a known and well-behaved situation.

A similar approach to adding noise in order to reduce the gap between training and test conditions was employed in the past by Van Compernelle and others [9] [10]. In these works, noise was synthetically introduced in the parameterized signal in the form of simple masking constants added to each frequency channel or as a random event generated using a Gaussian distribution model. On the contrary, in this paper we use real noises that are directly added to the signal in the time domain. We make use of different types of masking noises at different SNRs and show their relative performance in different tasks. In particular, we show how this approach is very intuitive and offers very significant improvements for the case of cross-talk noise: when the masking noise employed is babble noise (such as in a cafeteria), we show that the cross-talk speech is almost perfectly masked. It is also shown that when multiple recognizers employing different masking noises are available their outputs may be combined using ROVER [11] for increased accuracy rates. This is very important when the corrupting noises vary rapidly, although resulting in a more resource consuming process.

In Section 2 we introduce the strategy for improving robustness on noise-corrupted speech by the addition of well-behaved masking noises. In Section 3 we propose the combination of multiple masking noises for improved recognition using ROVER output transcription combination. Section 4 shows experimental results on a variety of corrupting noises and masking settings and we make conclusions in Section 5.

This work was done as part of an internship in IBM

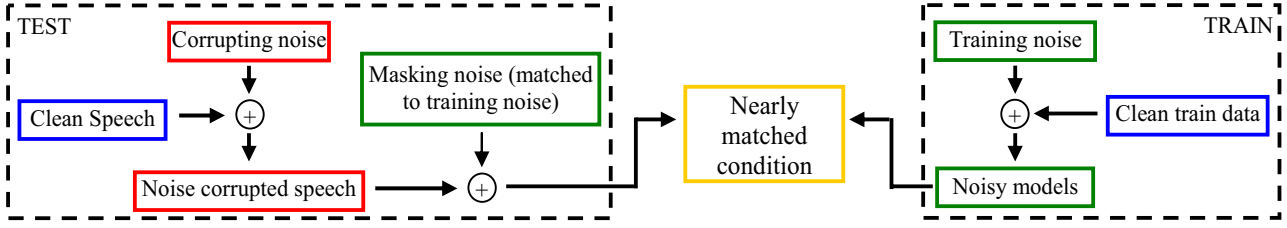


Figure 1: Flow diagram of the masking noise approach.

## 2. Robustness Increase by Adding a Masking Noise

### 2.1. Problem Statement

The method of masking noises is motivated by the perceptual observation that stationary (or quasi-stationary) noises may mask other noises in the human ear, to a certain extent. For example, the sound from a fan or even TV may be used to reduce the inconvenience caused by annoying neighbors. In this work, we use the same principle with the goal of turning an unknown and noise-variable environment into a known and more stable condition.

Fig. 2 shows four spectrograms derived from a common speech utterance. Subfigure a) shows the original clean speech file and b) the file corrupted by cross-talk noise (speech from a different speaker was added). A comparison between the two clearly shows the presence of the corrupting noise (more noticeable in the circled areas) that will cause a significant recognition accuracy loss. The lower figures are obtained by adding a masking noise (babble noise) to the previous spectrograms a) and b), obtaining c) and d), respectively. Comparison between c) and d) shows that the presence of the corrupting cross-talk noise is almost unnoticeable, indicating successful masking. Our hypothesis is that a system trained with clean speech like in a) and tested with noise-corrupted speech like in b) will be highly mismatched, while when the system is trained with speech plus masking noise as in c) and tested with noise-corrupted speech plus masking noise as in d) the mismatch will be significantly reduced.

### 2.2. Mathematical Formulation

From a computational point of view, adding a masking noise can be shown to reduce the gap between unmatched training and test conditions when logarithmic computation is involved. A noisy signal may be represented as:

$$y(t) = x(t) + n(t), \quad (1)$$

where  $x(t)$  is the clean speech and  $n(t)$  is a noise signal. Assuming statistical independence, the energy distribution in the frequency space is given by:

$$|Y(f)|^2 = |X(f)|^2 + |N(f)|^2, \quad (2)$$

and the expectation is:

$$E\{|Y(f)|^2\} = E\{|X(f)|^2\} + E\{|N(f)|^2\}. \quad (3)$$

Similarly, if a masking noise is added, the expectation is:

$$E\{|Y_M(f)|^2\} = E\{|X(f)|^2\} + E\{|N(f)|^2\} + E\{|M(f)|^2\}, \quad (4)$$

where  $M(f)$  stands for the masking noise. Suppose that two acoustic model sets are trained for clean, and clean with masking noise distributions, respectively:

$$E\{|A(f)|^2\} = E\{|X(f)|^2\}, \quad (5)$$

and

$$E\{|A_M(f)|^2\} = E\{|X(f)|^2\} + E\{|M(f)|^2\}, \quad (6)$$

where we use  $A(f)$  to refer to the acoustic model's spectral representation trained with clean speech and  $A_M(f)$  for the models trained with speech with masking noise.

Now, dropping the expectation symbol for ease of notation, the difference between the distributions of the acoustic models and data in the logarithmic space is, in each case:

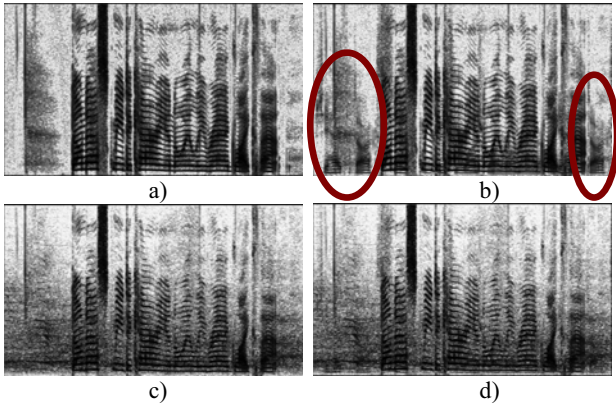
$$\Delta_1 = \log |Y(f)|^2 - \log |A(f)|^2 = \log \left( 1 + \frac{|N(f)|^2}{|X(f)|^2} \right) \quad (7)$$

$$\begin{aligned} \Delta_2 &= \log |Y_M(f)|^2 - \log |A_M(f)|^2 = \\ &= \log \left( 1 + \frac{|N(f)|^2}{|X(f)|^2 + |M(f)|^2} \right). \end{aligned} \quad (8)$$

From Eq. (7) and (8), the gap between the expectations of model and test data is reduced when the masking noise increases. However, adding noise reduces the global SNR (and ASR accuracy). Thus, a tradeoff is needed between the ability to mask corrupting noises and the reduction of SNR.

In this work we make emphasis on a particularly complicated type of distortion, cross-talk noise, and on the use of babble noise (background speech from multiple sources) to mask it. The *central limit theorem* states that given a set of independent random variables, their sum approximates a Gaussian distribution as long as the number of sources is sufficiently large. Also, for large numbers of variables the addition or removal of any particular source does not affect the result of the theorem. Now we apply this to our particular masking noise, babble noise that is a collection of speakers talking simultaneously. Although an individual speaker's signal is a highly non-stationary process, when several of them are combined the result is an almost-stationary noise. Since cross-talk noise is also produced by another speaker, we expect babble noise to be able to absorb the corrupting speaker within the quasi-stationary process, when it is added at an appropriate power level.

In the future, we will investigate the possibility of using the same principle to create other types of masking noises adequate for masking particular corrupting noises. For example, for the noise observed on-board cars, a masking noise could be created as the combination of noises collected in mobile environments and employed to absorb the corrupting noise.



**Figure 2:** Spectrograms derived from *DR1\_FAKS0\_SAI* in TIMIT. a) Original file. b) Original file corrupted with cross-talk noise (15dB). c) Original file plus masking noise. d) Original file plus cross-talk noise and masking noise. Red ovals in b) show regions where the presence of corrupting noise is more evident.

### 3. ROVER Combination of Output Transcriptions

The reliability of a particular masking noise at a fixed energy level is subject to variations because the corrupting noises affecting speech will typically vary in energy and spectral shape. Also, when masking noises are inserted there is a tradeoff between mismatch reduction and global SNR decrease. Thus, in order to optimize performance different masking noises may be used for different portions of speech.

In this work, outputs from recognizers using different masking noises at different SNRs were combined using a ROVER strategy [11] to provide increased accuracy. An advantage of the masking noise strategy is that a large number of recognizer outputs can be generated modifying the corrupting noise but keeping the same architecture and training methodology.

### 4. Results and Discussion

Performance is evaluated in terms of phonetic accuracy recognition according to:

$$\%acc = \frac{corr - ins}{corr + subs + del} \quad (9)$$

The prototype of phonetic recognition engine used consists of 51 Hidden Markov Models (HMM) and a phone bigram. Models are left-to-right with 3 emitting states each, and 15 Gaussians per state. The front-end uses pre-emphasis ( $\alpha=0.97$ ) and 25ms Hamming windows with a 10ms window shift. Thirteen MFCC coefficients including C0 and their respective first and second order derivatives (39 total features) are computed from a filter-bank of 26 Mel-scaled filters distributed in the region 0-8 kHz and Cepstral Mean Normalization was used. In addition to the base system, other recognizers were built with the same characteristics, only adding different training noises to the training data. Training and test data are the male speaker files in the training and test partitions of TIMIT, respectively. Training, masking and corrupting noises are artificially added to the clean TIMIT files, where applies.

Corrupting Noise → TrainNse – MaskNse ↓	Battle Noise SNR: 5dB	Cross-talk Noise SNR: 15dB	Humvee vehicle SNR: 5dB
Baseline	56.48	56.14	53.65
White - noMask	57.51 (2.4)	56.49 (0.8)	57.69 (8.7)
<b>White - White</b>	<b>58.61 (4.9)</b>	<b>59.74 (8.2)</b>	<b>60.25 (14.2)</b>
Babble - noMask	58.21 (4.0)	56.40 (0.6)	61.97 (18.0)
<b>Babble - Babble</b>	<b>59.72 (7.4)</b>	<b>63.69 (17.2)</b>	<b>62.49 (19.1)</b>

**Table1.** Absolute percent accuracy with and without masking (in parentheses relative increase compared to the baseline). Accuracy of a matched clean system is 72.43%.

#### 4.1. Performance Increase Adding Masking Noises

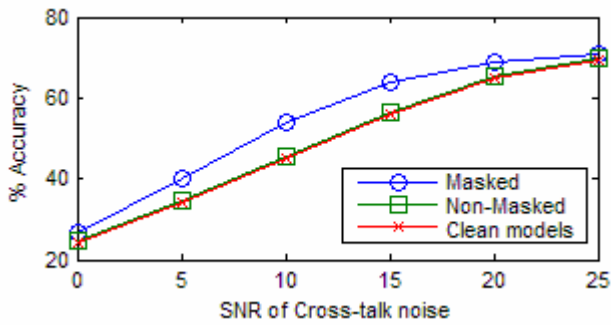
When the corrupting noise or SNR level in a speech signal is unknown, but an approximate idea of its behavior is available, a generic training noise may be added to the training data in order to generate noisy speech models that better match noise-corrupted speech. However, performance will always be degraded compared to the case of perfect match between training and test conditions. We propose to add to the test data a masking noise of the same type as that employed for training the models. With this, we expect to mask the corrupting noise and reduce the mismatch between training and test. Table 1 shows performance for a variety of combinations of corrupting and training noises. Results are given in absolute values, as well as relative to the baseline accuracy (baseline is ASR of noisy speech with clean models). Rows are paired to show in each case performance with and without the use of masking in the test data.

Noisy models improve ASR of noise-corrupted speech in every situation. The most interesting observation is that adding a masking noise matched to the training noise significantly improves performance, especially in the case of cross-talk noise: using babble noise for masking cross-talk corrupting speech produced a very promising relative improvement of 17.2% over the baseline accuracy, while only 0.6% was obtained by using training noise and not masking noise. In Fig. 3, results are given for different SNR levels of cross-talk corrupting speech. Plots show that clean models and noisy models trained with babble noise produce almost identical results, while the use of a masking noise significantly improves performance.

An important parameter required in our approach is the optimal SNR level of training noise and masking noise. Training and masking noises should be matched so that training and test are done under equivalent conditions. Also, the amount of noise added should be enough to mask the corrupting noise, but should be kept minimal to avoid significant global SNR decrease. In this work, accuracy results correspond to the noise levels that maximize performance, though in real applications automatic noise analysis should be employed to find the appropriate noise energy levels. Also, constant levels were used for the entire length of each speech file. However, in future work we expect to improve results by applying different levels of masking noise according to the power level of the corrupting noise in each speech segment.

#### 4.2. Combination of Output Transcriptions from Different Masking Noises

So far, we considered performance using models trained with a single type of noise. However, the combination of different training noises may improve performance.



**Figure 3:** Accuracy for cross-talk corrupting noise at different SNRs. Results are given for clean models and models trained with babble noise, with and without masking noise added to test data.

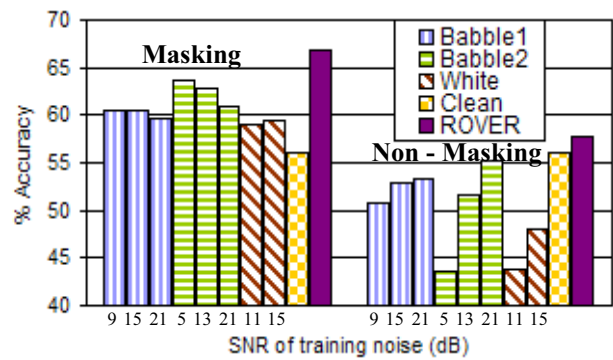
A simple and efficient technique is multi-style training, consisting of training new models with data from multiple noise conditions. However, the masking noise approach cannot benefit from multi-style training, because training and masking noises should be matched, and two different noise types cannot be matched at the same time. We propose an alternative consisting of the use of multiple masking-noise-based recognizers and the combination of their outputs by means of ROVER voting. Fig. 4 shows results for the case of cross-talk corrupting noise and compares performance with and without masking noise. Each bar represents an individual recognizer's accuracy (each pattern represents a type of training noise and different SNR levels were also used in each case), bars filled with squares are for recognition with clean models and the solid bars show the combined accuracy using ROVER. Individual recognizers using masking noise show a very robust behavior, compared to those when no masking noise is used. As the individual components perform significantly better when masking is used, so does the combined result (a relative accuracy gain of 8.7% is obtained over the best individual output, as opposed to only 3.5% when no masking noise is used). The relative accuracy gain over the system trained with clean speech (the solid bars) is 24.2% when masking is used and only 3.5% when it is disabled. For comparison, accuracy was evaluated for a multi-style recognizer trained with data from the same noise types and SNRs but it did not even outperform the system trained with clean speech, which shows the difficulty of the cross-talk noise task for conventional approaches.

## 5. Conclusions

In this work we presented a robust recognition scheme where a stationary noise is used to mask the negative effects of unknown noises in speech recognition. The same type of noise is added to both the training and test data in a speech recognizer, reducing the mismatch. Noise is a random process and thus, in order to tackle it another random process is more appropriate than a deterministic one.

The masking noise approach improved accuracy in every situation considered. A particularly interesting case is cross-talk speech masked with babble noise. If the SNR between target speaker and noise speaker is sufficiently high, the later may be absorbed in the amalgam of background speakers that compose babble noise. In the future, the same principle could be used to design combinations of noise for optimal masking of particular noise types.

In this work the term *well-behaved* was used referring to masking noises, and this was translated in practice into the quasi-stationarity of noise. However, more work is intended to gain insights on the most appropriate types of masking



**Figure 4:** Accuracy of individual recognizers and their combination using ROVER for cross-talk corrupting noise at 15dB. Results with and without masking are shown.

noises and other constraints. With the same aim a comparison will also be made with the method proposed by van Compernelle, where synthetic noise was employed. Also, in the future we expect to increase performance by adding different energy levels of masking noise according to the energy level of the corrupting noise in the speech signal.

## 6. References

- [1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech, and Signal Processing, IEEE Trans.*, vol. 28 (4), pp. 357-366, August, 1980.
- [2] H. Hermansky and N. Morgan, "RASTA processing of speech," *Speech, and Audio Processing, IEEE Trans.*, vol. 2 (4), pp. 578-589, October, 1994.
- [3] J. A. Nolasco-Flores and S. J. Young, "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation," *Proc. ICASSP'94*, pp. 409-412, April, 1994.
- [4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Jour. Acoustical Society of America*, vol. 55, pp. 1304-1312, June, 1974.
- [5] J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," *Proc. Eurospeech'01*, pp. 217-220, September, 2001.
- [6] C. J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer, Speech and Language*, vol. 9 (2), pp. 171-185, April, 1995.
- [7] J. L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Speech and Audio Processing, IEEE Trans.*, vol. 2 (2), pp. 291-298, April, 1994.
- [8] J. Droppo, A. Acero and Li Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," *Proc. ICASSP'02*, pp. 57-60, May, 2002.
- [9] D. Van Compernelle, "Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction", *Proc. ICASSP'87*, pp. 1143-1146, April, 1987.
- [10] T. Claes and D. Van Compernelle, "SNR-normalisation for robust speech recognition", *Proc. ICASSP'96*, pp. 331-334, May, 1996.
- [11] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," *Proc. IEEE ASRU'97*, pp. 347-354, December, 1997.