# Robust Phonetic Feature Extraction Under a Wide Range of Noise Backgrounds and Signal-to-Noise Ratios

*Shuangyu Chang, Lokendra Shastri and Steven Greenberg*

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
{shawnc; shastri; steveng@icsi.berkeley.edu}

## Abstract

A method for automatic classification of articulatory-acoustic features (AFs) and phonetic segments has been developed that is relatively immune to performance degradation under a wide range of acoustic-interference conditions. A key property of the classification method is to train on two separate noise backgrounds ("pink" and "white" noise) across a 30-dB dynamic range of signal-to-noise ratios (SNRs). This training regime reduces the error rate at the articulatory-feature and phonetic-segment levels by as much as 40-60% for low-SNR conditions relative to the baseline system (trained solely on "clean" speech) and thus ensures that phonetic-segment classification is sufficiently high (60-80% accuracy) as to provide reasonably robust word recognition performance at low SNRs.

## 1. Introduction

Acoustic interference poses a significant challenge to current-generation automatic speech-recognition (ASR) systems. ASR systems that work well under pristine acoustic conditions generally perform much more poorly at low signal-to-noise ratios (SNRs). In contrast, human listeners typically experience little (if any) degradation of intelligibility under comparable circumstances, except for SNRs of less than 0 dB [10]. The robust nature of human speech decoding may reflect the brain's application of multiple processing strategies, spanning a broad range of time constants and structural units, providing complementary perspectives on the signal's phonetic and lexical representation [5][6]. Motivated by such concerns, we have developed a method for decoding the speech signal into articulatory-acoustic features (AFs) and phonetic segments across a wide dynamic range of acoustic background conditions that is relatively robust to variation in signal-to-noise ratio and spectro-temporal character of the acoustic background.

The current study focuses on classification of articulatory-acoustic features, as previous research has demonstrated that these "atomistic" units of the speech signal are more robust to acoustic interference than phonetic segments [9]. Phonetic segments are subsequently derived from clusters of AFs, using MLP networks similar to those employed in a previous study of phonetic classification [2]. Under such conditions the reduction in error rate relative to the baseline system is often far better than would be expected using a conventional phonetic-segment classification system (as employed in [4]).

A singular property of the current method is that the system is trained not only on "clean" speech (i.e., speech that has been recorded under pristine, high-SNR, conditions), but also on material embedded in a variety of noise backgrounds over a wide dynamic range of SNRs. This training regime is motivated by the supposition that the robustness of human speech recognition is likely to derive from exposure to a broad range of acoustic interference conditions and reflects the ability of the brain to generalize from specific patterns of acoustic interference to novel noise conditions.

## 2. Corpus Materials

The speech material used in the current study is derived from the Numbers95 corpus, collected and phonetically annotated (i.e., labeled and segmented) at the Oregon Graduate Institute [3]. This corpus contains the numerical portion (mostly street addresses and phone numbers) of thousands of telephone dialogues and possesses a lexicon of 37 words and an inventory of 29 phonetic segments. The speakers contained in the corpus are of both genders, and represent a wide range of dialect regions and age groups. The speech material was recorded with 16-bit resolution at an 8-kHz sample rate. The training set for the baseline system contains ca. 2.5 hours of material. A separate 15-minute, cross-validation set was used during training to minimize the chances of overfitting the corpus data. Testing was performed on an independent set of material (of ca. 60-minutes' duration) from the same corpus. Various forms of acoustic interference, derived from the NOISEX corpus [12], were mixed (in additive fashion) with the OGI Numbers95 speech material. The NOISEX material was originally recorded with 16-bit resolution at 19.98 kHz but was down-sampled to 8 kHz for the current study. A subset of the noise backgrounds was mixed with the speech material over a range of SNRs (as indicated in Table 1). The signal-to-noise ratio was calculated from the normalized power (computed over the entire length of the utterance) for both the speech signal and the noise background using a procedure described in [8].

## 3. Baseline-System Processing and Training

The speech signal was processed in several stages for the baseline system. First, a power spectrum was computed every 10 ms (over a 25-ms window) and this spectrum partitioned into quarter-octave channels between 0.3 and 4 kHz. The power spectrum was logarithmically compressed in order to preserve the general contour of the spectrum distributed across frequency and time. An array of independent, multilayer perceptron (MLP) networks classified each 25-ms frame along five articulatory-based, phonetic-feature dimensions: (1) place (e.g., labial, velar, alveolar) and (2) manner of articulation (e.g., stop, fricative, nasal, vocalic), (3) voicing (voiced/unvoiced), (4) lip-rounding (rounded/unrounded) and (5) front-back articulation (for vocalic segments). A separate feature was trained for "silence" (labeled as "null" in each feature dimension). These articulatory-feature labels were subsequently combined and served as input to a separate MLP that performed classification of phonetic-segment identity (cf. [2] for additional information regarding the structure of the AF classifiers and the combination of their output for phone classification). The baseline system was trained only on "clean" speech while testing was performed on clean, as well as on a variety of noisy speech material at different SNRs.

## 4. Training Over a Wide Range of Noise Backgrounds

A training regime was developed for enhancing baseline system performance via utilization of an assortment of noise back-

| | | Manner | | | Place | | | Front-Back | | | Voicing | | | Lip-Rounding | | | Phones | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Noise/Condition** | **SNR** | **Base** | **Mix** | **E-R** | **Base** | **Mix** | **E-R** | **Base** | **Mix** | **E-R** | **Base** | **Mix** | **E-R** | **Base** | **Mix** | **E-R** | **Base** | **Mix** | **E-R** |
| **Clean** | — | 82.6 | 81.8 | -4.6 | 75.0 | 76.4 | 5.6 | 82.6 | 82.8 | 1.1 | 89.8 | 89.5 | -2.9 | 83.4 | 83.3 | -0.6 | 78.1 | 79.2 | 5.0 |
| **Pink** | 0 | 30.0 | 66.2 | 51.7 | 32.8 | 61.9 | 43.3 | 46.9 | 68.6 | 40.9 | 53.3 | 77.1 | 51.0 | 48.3 | 69.4 | 40.8 | 21.9 | 64.1 | 54.0 |
| | 10 | 55.7 | 76.3 | 46.5 | 54.0 | 71.7 | 38.5 | 63.5 | 78.4 | 40.8 | 74.2 | 85.0 | 41.9 | 64.5 | 79.1 | 41.1 | 55.2 | 74.6 | 43.3 |
| | 20 | 76.1 | 81.2 | 21.3 | 71.5 | 76.3 | 16.8 | 78.9 | 82.4 | 16.6 | 86.2 | 88.8 | 18.8 | 79.8 | 83.2 | 16.8 | 72.9 | 79.0 | 22.5 |
| | 30 | 81.8 | 82.3 | 2.7 | 74.7 | 77.0 | 9.1 | 81.9 | 83.3 | 7.7 | 89.0 | 89.7 | 6.4 | 82.9 | 83.9 | 5.8 | 77.3 | 79.7 | 10.6 |
| **White** | 0 | 28.4 | 65.2 | 51.4 | 30.5 | 60.3 | 42.9 | 44.6 | 66.6 | 39.7 | 57.3 | 77.5 | 47.3 | 45.5 | 67.5 | 40.4 | 18.9 | 61.2 | 52.2 |
| | 10 | 48.1 | 74.0 | 49.9 | 46.8 | 69.8 | 43.2 | 56.7 | 76.1 | 44.8 | 69.9 | 84.2 | 47.5 | 57.7 | 76.9 | 45.4 | 45.2 | 72.4 | 49.6 |
| | 20 | 67.6 | 79.9 | 38.0 | 65.7 | 75.0 | 27.1 | 73.0 | 81.2 | 30.4 | 82.1 | 88.1 | 33.5 | 73.2 | 82.1 | 33.2 | 67.5 | 77.8 | 31.7 |
| | 30 | 80.0 | 82.0 | 10.0 | 73.6 | 76.8 | 12.1 | 81.1 | 83.0 | 10.1 | 88.7 | 89.5 | 7.1 | 81.8 | 83.9 | 11.5 | 76.1 | 79.5 | 14.2 |
| **Mixture of White and Pink Noise*** | 0 | 29.3 | 65.6 | 51.3 | 32.2 | 61.2 | 42.8 | 46.3 | 67.8 | 40.0 | 53.6 | 76.7 | 49.8 | 47.6 | 68.7 | 40.3 | 20.4 | 63.2 | 53.8 |
| | 10 | 52.9 | 75.5 | 48.0 | 51.5 | 71.2 | 40.6 | 61.1 | 77.8 | 42.9 | 72.2 | 84.7 | 45.0 | 62.2 | 78.4 | 42.9 | 52.0 | 74.0 | 45.8 |
| | 20 | 74.0 | 81.0 | 26.9 | 70.3 | 76.0 | 19.2 | 77.6 | 82.2 | 20.5 | 85.4 | 88.6 | 21.9 | 78.4 | 83.0 | 21.3 | 71.7 | 78.7 | 24.7 |
| | 30 | 81.5 | 82.3 | 4.3 | 74.6 | 77.0 | 9.4 | 81.8 | 83.2 | 7.7 | 88.9 | 89.7 | 7.2 | 82.8 | 84.0 | 7.1 | 77.1 | 79.7 | 11.4 |
| **Speech Babble*** | 0 | 33.4 | 51.7 | 27.5 | 35.2 | 46.6 | 17.6 | 48.4 | 57.3 | 17.2 | 50.0 | 63.5 | 27.0 | 49.1 | 58.4 | 18.3 | 28.4 | 45.2 | 23.5 |
| | 10 | 62.2 | 70.4 | 21.7 | 58.7 | 64.9 | 15.0 | 67.6 | 71.9 | 13.3 | 74.8 | 77.5 | 10.7 | 68.3 | 73.0 | 14.8 | 57.8 | 68.3 | 24.9 |
| **Buccaneer Jet Cockpit (190 knots)*** | 0 | 28.5 | 63.3 | 48.7 | 30.8 | 57.6 | 38.7 | 45.9 | 64.6 | 34.6 | 52.2 | 74.7 | 47.1 | 47.2 | 66.0 | 35.6 | 19.7 | 59.2 | 49.2 |
| **Buccaneer Jet Cockpit (450 knots)*** | 0 | 29.5 | 61.2 | 45.0 | 33.1 | 57.0 | 35.7 | 46.6 | 65.0 | 34.5 | 52.5 | 72.1 | 41.3 | 48.2 | 65.6 | 33.6 | 20.6 | 58.5 | 47.7 |
| **F-16 Jet Fighter Cockpit*** | 0 | 27.9 | 62.7 | 48.3 | 30.4 | 57.2 | 38.5 | 45.3 | 65.0 | 36.0 | 50.1 | 74.4 | 48.7 | 46.5 | 66.3 | 37.0 | 23.5 | 58.8 | 46.1 |
| **Destroyer Engine Room*** | 0 | 26.0 | 59.1 | 44.7 | 29.2 | 52.8 | 33.3 | 43.7 | 59.6 | 28.2 | 52.3 | 72.9 | 43.2 | 44.8 | 62.0 | 31.2 | 20.8 | 51.1 | 38.3 |
| **Destroyer Operations Room*** | 0 | 37.2 | 59.1 | 34.9 | 38.1 | 53.9 | 25.5 | 51.1 | 64.0 | 26.4 | 55.6 | 69.4 | 31.1 | 52.4 | 64.9 | 26.3 | 34.3 | 58.0 | 36.1 |
| **Leopard 2 Military Vehicle*** | 0 | 61.5 | 64.3 | 7.3 | 57.0 | 58.6 | 3.7 | 66.7 | 66.3 | -1.2 | 69.0 | 69.7 | 2.3 | 66.2 | 68.3 | 6.2 | 56.5 | 62.7 | 14.3 |
| **M109 Tank*** | 0 | 46.1 | 67.6 | 39.9 | 45.5 | 62.6 | 31.4 | 56.4 | 69.4 | 29.8 | 63.8 | 76.5 | 35.1 | 57.1 | 71.0 | 32.4 | 43.0 | 66.1 | 40.5 |
| **Machine Gun*** | 0 | 65.3 | 66.9 | 4.6 | 59.9 | 61.7 | 4.5 | 70.7 | 70.7 | 0.0 | 73.8 | 75.2 | 5.3 | 71.6 | 71.7 | 0.4 | 59.3 | 63.4 | 10.1 |
| | 10 | 72.7 | 72.6 | -0.4 | 65.9 | 66.7 | 2.3 | 75.4 | 75.0 | -1.6 | 79.6 | 80.0 | 2.0 | 76.1 | 75.6 | -2.1 | 67.9 | 70.6 | 8.4 |
| **Car Factory (Floor)*** | 0 | 30.8 | 57.9 | 39.2 | 33.5 | 54.0 | 30.8 | 47.3 | 62.8 | 29.4 | 52.5 | 69.4 | 35.6 | 48.3 | 63.6 | 29.6 | 25.1 | 52.5 | 36.6 |
| | 10 | 59.2 | 73.0 | 33.8 | 56.7 | 68.2 | 26.6 | 65.9 | 75.2 | 27.3 | 74.7 | 81.5 | 26.9 | 67.0 | 76.0 | 27.3 | 55.4 | 70.6 | 34.1 |
| **Car Factory (Production Hall)*** | 0 | 41.4 | 68.3 | 45.9 | 42.0 | 63.2 | 36.6 | 53.8 | 69.9 | 34.8 | 61.5 | 77.5 | 41.6 | 54.4 | 71.8 | 38.2 | 37.6 | 65.7 | 45.0 |
| **Volvo (Passenger Compartment)*** | 0 | 76.7 | 78.4 | 7.3 | 70.0 | 71.3 | 4.3 | 78.4 | 78.1 | -1.4 | 83.9 | 84.3 | 2.5 | 79.2 | 78.9 | -1.2 | 70.4 | 74.8 | 14.9 |
| | 10 | 80.3 | 79.7 | -3.0 | 72.5 | 73.2 | 2.5 | 80.4 | 80.5 | 0.5 | 87.6 | 87.7 | 0.8 | 81.4 | 81.3 | -0.6 | 74.6 | 76.9 | 9.1 |
| **High-Frequency Radio Channel*** | 0 | 26.5 | 64.0 | 51.0 | 29.0 | 58.1 | 41.0 | 43.6 | 64.0 | 36.2 | 57.4 | 77.1 | 46.2 | 44.4 | 65.4 | 37.8 | 17.9 | 58.0 | 48.8 |

**Table 1** Classification accuracy at the articulatory-acoustic-feature and phonetic-segment levels as a function of noise-background condition and signal-to-noise ratio (SNR, in dB). Background noises are from the NOISEX corpus. Conditions marked with an asterisk (*) are those for which the system was tested but not previously trained (i.e., conditions "unseen" by the classifiers prior to testing). For each articulatory-feature dimension (place, front-back, manner, voicing, lip-rounding) and phonetic segment ("phones") classification experiment, recognition accuracy is shown in terms of percent correct for the baseline ("Base") and enhanced, noise-trained ("Mix") system. The percent reduction of error ("E-R") yielded by the "Mix" system relative to the "Base" system is shown to the right of classification performance (and is marked in blue).

grounds and SNRs. The AF and phone-segment classification system was trained on speech material embedded in both white and pink noise over a 30-dB range of SNRs (as well as on "clean" speech, as described in Section 3) in order to ascertain the degree of transfer to conditions "unseen" during training (marked by an asterisk, *, in Table 1).

## 5. Baseline Performance

Baseline-system performance under "clean" conditions yields ca. 78% correct phonetic-segment classification (consistent with ca. 5% error rate at the word level - cf. [11] for details). Articulatory-feature classification for this "clean" condition ranges between 75% and 90% correct, depending on the AF dimension involved (cf. Table 1). In circumstances where the speech material is embedded in background noise the baseline system performs well primarily when the SNR is 20 dB or higher. For lower SNRs classification performance is often less than 60% correct. At the lowest SNR (0 dB), performance often degrades to below 40% correct for articulatory place and manner classification. Phonetic-segment classification performance for these conditions often falls below 30% correct. The pattern of degradation as a function of SNR is clearly manifest in Figure 1 for three forms of background noise ("pink", "white" and a "training-blind" pink/white "hybrid" whose spectral slope is midway between that of pink and white noise). The AF dimensions of voicing and rounding, both binary in nature, outperform (by a slight degree) the other AF dimensions. The place dimension is classified with the lowest accuracy of the articulatory-feature dimensions, consistent with other studies of automatic AF classification (e.g., [1][13]).

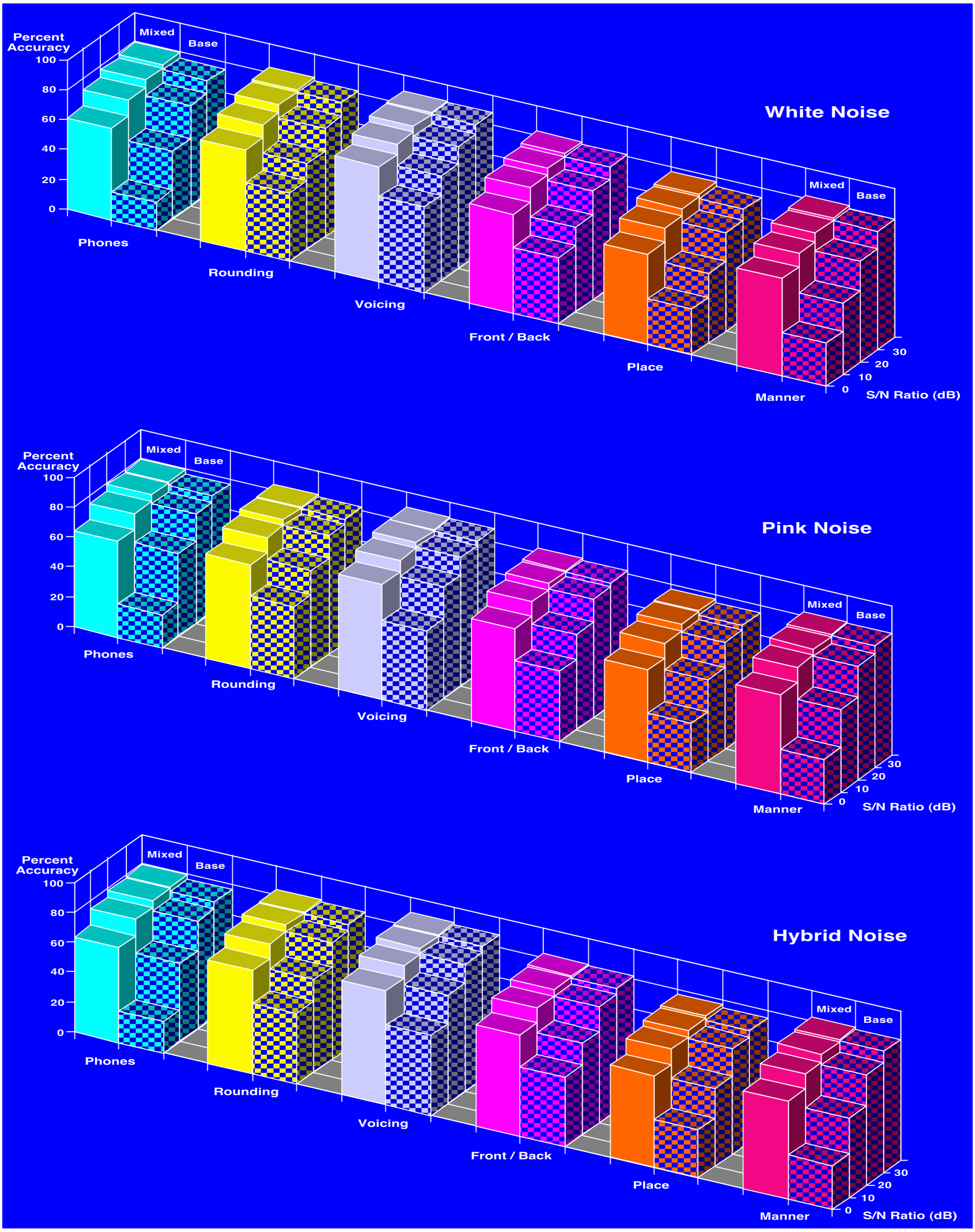

**Figure 1** Articulatory-acoustic-feature and phonetic-segment classification accuracy of the baseline (“Base”) and mixed-training (“Mixed”) systems as a function of signal-to-noise ratio for three different noise backgrounds (white, pink and a novel, untrained, hybrid noise). Note, that in all instances the mixed-training system yields classification performance that degrades relatively little as a function of decreasing signal-to-noise ratio. In contrast, the baseline system exhibits significant degradation of classification accuracy as the SNR drops below 20 dB. Data are a subset of the complete classification results described in Table 1.

## 6. Mixed-Training System Performance

In contrast to the baseline system, performance for the mixed-training system degrades relatively little as a function of SNR (cf. Figure 1 for performance in pink, white and hybrid noise). Classification accuracy is typically better than 60% for most of the background-noise conditions, even at the lowest SNRs. What is particularly noteworthy is the ability of the mixed-training system to generalize from training purely on “clean” speech, as well as pink and white noise, to acoustic backgrounds with very different spectro-temporal properties (these conditions are marked with an asterisk,*, in Table 1). In virtually all instances the mixed-trained system outperforms the baseline system, often by 20-50% (the reduction in error rate, “E-R,” is marked in blue in Table 1). The largest gain in error reduction (relative to the baseline system) is generally observed for conditions associated with the lowest SNRs (0 and 10 dB). For these conditions an error-rate reduction of 40% or greater is not uncommon. The conditions where the mixed-training system fails to significantly outperform the baseline system are those in which the latter is already performing close to the maximum associated with the classification framework used. In such instances there is generally little difference in classification accuracy between the baseline and mixed-training systems.

## 7. Implications for Robust Speech Recognition

Acoustic interference is currently the bane of automatic speech recognition systems. Currently, the most effective means of coping with noise of variable (and unforeseen) character and magnitude is for the speaker to use a close-fitting microphone that effectively filters out background interference. Yet, there are many circumstances where special-purpose microphones are unavailable and where reliance on noise-robust speech recognition algorithms is necessary. In the past, most noise-robust ASR algorithms relied on training using acoustic-background materials similar to those likely to be encountered in the task (i.e., test) domain. The current study demonstrates that training over a circumscribed set of noise backgrounds (white and pink) encompassing a broad (30-dB) range of SNRs is capable of imparting a significant degree of generality to the noise-background training sufficient to yield good recognition performance across a wide range of acoustic conditions.

It is likely that incorporation of additional linguistic features at levels superordinate to that of the articulatory feature and phonetic segment would yield an even greater reduction in error rate than afforded by the current set of feature dimensions (cf. [5] for discussion of this issue).

Although the current study does not directly address the issue of word-level recognition in noise backgrounds, the results are also likely to be of significance for word recognition performance because of the high degree of correlation between word-recognition performance and accuracy of phonetic-segment classification (cf. [7] for a discussion of this relation). Therefore, good performance on phonetic-segment classification across a broad range of acoustic backgrounds is likely to yield a relatively high level of word-recognition accuracy under comparable conditions.

## 8. Acknowledgements

The research described in this paper was supported by the Learning and Intelligent Systems Initiative of the National Science Foundation.

## References

[1] Chang, S., Greenberg, S. and Wester, M. “An elitist approach to articulatory-acoustic feature extraction,” *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001),* 2001.

[2] Chang, S., Shastri, L and Greenberg, S. “Automatic phonetic transcription of spontaneous speech (American English),” *Proceedings of the International Conference on Spoken Language Processing*, pp. 330-333, 2000.

[3] Cole, R., Fanty, M., Noel, M. and Lander, T. “Telephone speech corpus development at CSLU,” *Proceedings of the International Conference on Spoken Language Processing,* pp. 1815-1818, 1994.

[4] Furui, S. “Noise adaptation of HMMs using neural networks,” *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pp. 160-167, 2000.

[5] Greenberg, S.“On the origins of speech intelligibility in the real world,” *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 23-32, 1997.

[6] Greenberg, S. “Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation,” *Speech Communication* 29: 159-176, 1999.

[7] Greenberg, S. and Chang, S. “Linguistic dissection of switchboard-corpus automatic speech recognition systems,” *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pp. 195-202, 2000.

[8] Kingsbury, B. *Perceptually Inspired Signal-Processing Strategies for Robust Speech Recognition in Reverberant Environments,* Ph.D. Thesis, University of California, Berkeley.

[9] Kirchhoff, K. *Robust Speech Recognition Using Articulatory Information,* Ph.D. Thesis, University of Bielefeld, 1999.

[10] Lippmann, R. “Speech recognition by machines and humans,” *Speech Communication,* 22:1-15, 1997.

[11] Shire, M. *Discriminant Training of Front-End and Acoustic Modeling Stages to Heterogeneous Acoustic Environments for Multi-Stream Automatic Speech Recognition*, Ph.D. Thesis, University of California, Berkeley, 2000.

[12] Varga, A., Steeneken, H., Tomlinson, M. and Jones, D. “The NOISEX-92 study on the effect of additive noise on automatic speech recognition.” DRA Speech Research Unit Technical Report, 1992.

[13] Wester, M., Greenberg, S. and Chang, S. “A Dutch treatment of an elitist approach to articulatory-acoustic feature extraction,” *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001),* 2001.