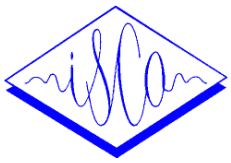

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

NOISE REDUCTION FOR SPEECH RECOGNITION AND SPEAKER VERIFICATION IN MOBILE TELEPHONY

Vijay Raman and Jayant Naik

NYNEX Science and Technology, Inc.
500 Westchester Avenue
White Plains, NY 10604 USA

ABSTRACT

Noise classification and reduction is used in this work to improve speech recognition and speaker verification performance for utterances received over the mobile network. Automatic noise classification is achieved with high accuracy in conjunction with spectral subtraction methods. The results audibly improve signal-to-noise ratio and significantly improve recognizer performance in low SNR situations as well as poorly modeled environments. The process is fully automatic, from implementation considerations.

I. INTRODUCTION

Speech recognition performance has improved considerably in recent years, but robustness against noise remains a significant problem in any real system. Many approaches have been proposed in the literature, with varying degrees of complexity([1]-[7],etc.). The specific environment considered in this work is that of *mobile (cellular) telephony* (e.g. [4]-[6]). Background noise can be extremely high, resulting in negative signal-to-noise ratios, in addition to the presence of line noise. Additionally, in the now very common *hands-free* environment, speech detection and squelching in the end-user equipment adds to the difficulty of noise reduction.

Noise reduction as part of the front end, is evaluated in this work, for speaker-dependent speech recognition and speaker verification of connected digits. A *spectral subtraction* algorithm is used, in conjunction with an *automatic noise-classification* algorithm for rapidly detecting and distinguishing noise segments. Spectral subtraction ([1]) is useful in recognition applications, but does require noise estimation, which, in this work, is performed in the classifier. The real-time noise estimation process is shown to perform reliably on a wide range of speech data. The noise reduction system is treated as independent of the speech recognition system.

The speech database used for testing was collected over the mobile network, and encompasses a wide range of speakers and calling conditions. A subset of the database, consisting of digits and names, was used in these evaluations.

Speaker-Dependent speech recognition and *Speaker Verification* performance was tested with and without noise reduction in the front end. Both DTW (Dynamic Time Warping) and HMM (Hidden Markov Model) systems were tested for speaker-dependent recognition, and as HMM system for speaker verification. A range of performance gains is demonstrated for the different systems.

II. MOBILE ENVIRONMENT / DATABASE

The speech database was collected over the mobile telephone network, from a population of 100 callers over a 6-month period, under a wide variety of calling conditions. The database comprises names, control words, isolated digits, and connected digit strings. A full description of this database is given in [8].

From review of the database, the following observations are made about speech acquired over the mobile network:

- the signal-to-noise-ratio can be extremely poor, negative in many cases,
- the noise sources, which include engine, tires, wind, and radio, result in noise components which are tonal and wideband, stationary and transient,
- within the duration of a session involving an isolated utterance, noise variation is generally not very significant.
- the use of hands-free (speaker-) phones introduces intra-session variation in the perceived background noise, described later.
- noise level and type variation from call to call can be very high.
- signal blanking is seen only with low frequency, for isolated utterances.

III. NOISE CLASSIFICATION

In the spectral subtraction approach to noise reduction, an estimate of the ambient noise is used to appropriately modify the signal. This method is dependent on obtaining a reasonably accurate estimate of the noise present during speech. The literature on recognition using this method may assume supervised estimate of such ambient noise, multi-channel input, or on-line estimation that may require extended signal observations.

Since this work required an automatic noise-reduction front-end, an algorithm for the identification and classification of appropriate noise segments for noise estimation was designed, which feeds into the spectral subtraction step. In designing the noise classifier, the following objectives were considered: (a) computational complexity was to be minimized, and (b) algorithm errors should be benign, i.e. errors in noise classification should distort the utterance minimally.

The classifier is primarily energy-based. Some additional related frame parameters are used to distinguish speech from noise. Segments of the signal are identified as possible representatives of ambient noise, and distinguished by level. This is essential, since different noise levels may be observed in various sections of the signal, seen at adjacent points in the transaction, particularly in the hands-free case. Algorithm parameters may be adjusted to make the algorithm more or less aggressive in seeking noise.

A significant aspect of the algorithm is that it may be operated in real-time, or in post-processing mode. In the post-processing mode, the classifier's decision is made after full utterance capture, and includes heuristics for deciding the appropriate noise segment to use. The post-processing mode provides improved performance in certain situations, the most important of which is the hands-free telephone, discussed below.

A problem occurs in noise estimation when the user has a hands-free phone - a situation increasingly common. The speech detection and squelch typically incorporated into these phones to prevent double-talk has the effect of "hiding" the ambient noise from the speech recognition system until speech actually begins. The true ambient noise is then "seen" only during the speech and for a short period after. In such a case, the post-processing mode of operation obtains the most appropriate estimate of the ambient noise. The post-processing mode can also be of value in cases where the speaker speaks immediately after the beep/prompt.

In terms of implementation, the post-processing mode is suitable for speech recognition of isolated utterances, in systems which are capable of utterance capture followed by fast analysis and recognition.

The classifier was found to work with an accuracy of about 85-90% compared to a human classifier. Errors are benign in most cases. The post-processing mode was used in the performance tests, noting that several of the users did have hands-free phones.

IV. NOISE REDUCTION

The signal segment identified by the classifier is utilized by a spectral subtraction algorithm in the front end. 256-sample frames with 50% overlap are used for FFT computation. Non-linear compression is used (e.g. [2], [6]), with non-linearity evidenced in low-level segments, eliminating tonal residuals.

In this implementation, the noise-reduction process is not tied to the recognizer, since it was to be compatible with other sub-systems consuming the received signal. The signal is reconstructed after noise-reduction, as in speech enhancement applications. Obviously, in a situation wherein the recognizer is the only consumer of noise-reduced data, reconstruction is not required, and feature extraction can proceed directly from the spectral representation of the noise-reduced signal.

Algorithm parameters were adjusted such that the effect of subtraction is reasonably benign even in cases where the classifier makes an error. Making the noise reduction more aggressive provides audibly cleaner speech in most cases, but is occasionally destructive to the utterance.

The post-processing mode is particularly helpful in the speaker-phone case. In such cases, real-time estimation does not provide sufficient noise reduction.

The effect of noise reduction on the energy-per-frame is shown in Figure 1. The energy histogram is computed over approximately 6000 utterances of original and noise-reduced speech data, including quiet sections. From this, it can be seen that during the low-energy segments, the signal energy has been reduced and redistributed, while the higher-energy segments, typically representing speech, have remained essentially unaffected. This is confirmed by listening tests.

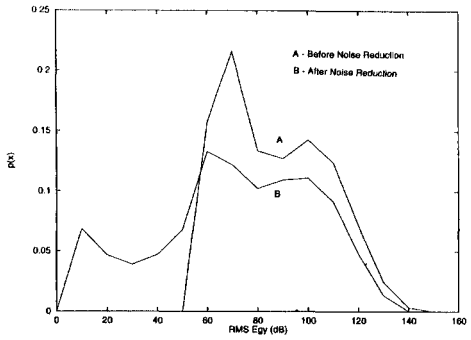

Fig. 1. Energy distribution for original and noise-reduced speech

V. SPEAKER-DEPENDENT RECOGNITION PERFORMANCE

Speaker-dependent recognition tests were performed on he mobile database described earlier. Tests on isolated digits and on names were performed on a round-robin basis, and the results are reported here.

Both training and testing tokens were noise-reduced. This assumes that noise-reduction is enabled at all times in the front end, both during enrollment and recognition.

In keeping with systems currently in operation, two training tokens were used for a speaker during a given test. Tests were conducted with training tokens selected on a round-robin basis over the full set of a speaker's utterances. It should be noted that the utterances were endpointed by hand for the purpose of training.

(While end-pointing is not the subject of this work, it appears that automatic endpointing will be significantly assisted by noise-reduction in the front-end.)

Initial evaluations of recognition performance were carried out on a DTW-based recognizer, on a subset of 4 speakers, 2 male and 2 female, with a vocabulary consisting of the isolated digits: 1-9, zero, oh. Each speaker generally had 10 or more repetitions of each vocabulary item. The results for this are reported in Table 1. The data was very noisy (mobile data) and it was found that very significant improvement in performance was achieved for the more noisy situations. In a quieter environment, the improvement is smaller.

Evaluations were then performed using the same digits data, using a HMM-based recognizer with a grammar for recognition of isolated utterances. The results are reported in Table 1. It was found during this work, that using background models built with data having undergone spectral subtraction improved performance, and this was therefore incorporated into the system.

Finally, a more extensive evaluation were performed using

the HMM recognizer with the isolated-utterance grammar above, with the data consisting of names instead of digits. There were 8 male and 6 female speakers in this test, and the names used by each speaker were specific to that speaker. Each speaker generally had 20 or more repetitions of a name. These results are reported in Table I.

(Note that all results are based on raw scores, without thresholding or other out-of-vocabulary rejection features active.)

TABLE I
SPEAKER-DEPENDENT RECOGNITION PERFORMANCE

Recognizer	Test Data	Orig. Error %	Noise Reduced Error %	Error Rate Reduction %
DTW	Digits	11.7	8.3	29
HMM	Digits	7.0	5.6	20
HMM	Names	3.1	1.2	61

It was useful to break down the results for the names data in the following way: the performance was evaluated for the four worst-performing speakers and the four best-performing speakers (before noise-reduction). These results are reported in Table II, and highlight a significant feature of the noise-reduction process: it generally provides the most improvement in high-noise situations, but *without degrading performance* in the low-noise situations. It was also found during the course of this work that in situations where the background is poorly modeled by the HMM background/silence models, the noise-reduction process improves performance.

TABLE II

Recognizer	Names Data - Subset	Orig. Error %.	Noise Reduced Error %	Error Rate Reduction %
HMM	4 worst	7.9	1.9	75
	4 best	0.6	0.6	0

VI. SPEAKER VERIFICATION PERFORMANCE

The impact of noise reduction on speaker verification performance in the cellular telephone network was evaluated using the noise-reduction algorithms described in Sections III and IV. A fixed-text verification protocol was used, and a ten-digit utterance such as a phone number was used as the voice password. Noise reduction was performed as a first step in the front-end processing. Training and testing were performed on both original and noise-reduced speech.

A subset of the speech database, comprising twenty speakers, each with eleven sessions, was chosen for evaluation of noise suppression and speaker verification. Each session consisted of three utterances of a ten-digit string for true-speaker tests, and three utterances of another ten-digit string which was shared with at least two other speakers, for imposter tests.

The speaker verification algorithm utilized LPC-Cepstral instantaneous and differential features, with continuous-density HMM models using state-specific full covariances. HMM digit-models were used with a fixed grammar. Speaker model adaptation and score normalization was also incorporated.

Noise reduction was applied to incoming speech and enrollment was performed by automatic segmentation of individual digits. Speaker-independent digits models built from noise-reduced speech data was used in this step to segment the digits. Subsequently, the speaker-specific digit models were used to perform speaker verification for each user. Three noise/background models were used along with the ten digit models.

A round-robin scheme was used by enrolling on speech data from five different sessions and performing verification on the remaining ten session for a total of fifty true-speaker sessions and one hundred imposter attempts for each of the twenty speakers.

The same experiment was performed on the original speech without noise-reduction.

The results are given in Table III.

TABLE III
SPEAKER VERIFICATION PERFORMANCE

	Orig %.	Noise Reduced %	Error rate reduction %
True-Speaker Rejection	1.0	1.0	-
Imposter Acceptance	16.5	11.2	30

VII. SUMMARY.

An automatic noise classification (estimation) and noise reduction system has been designed to work in the mobile telephony environment. The classifier, a key component of the system, performs well for isolated utterances in a range of noisy environments and adequately handles artifacts introduced in hands-free usage.

Evaluation of front-end noise reduction for speaker-dependent recognition indicates significant improvement in performance. Errors in high-noise errors are brought down considerably, without degradation under favorable conditions.Additionally, noise reduction appears to help compensate for HMM background/silence models that do not adequately model the actual background.

Evaluation of noise reduction in the context of a fixed-text speaker-verification system shows an improvement in speaker discrimination of over 30%. Accurate segmentation and more stable feature representation accrue from the front-

end noise reduction.

The utility of front-end noise reduction in improving the accuracy of endpointing isolated utterances (e.g. during training) requires further study.

REFERENCES

[1] Boll, S. "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27, 113-120 , 1979.

[2] Campernolle, D.V., "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, 3, 151-167, 1989.

[3] Erell, A. and Weintraub, M., "Energy Conditioned Spectral Estimation for Recognition of Noisy Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 1, No. 1, 84-89, 1993.

[4] Nakamura, S., Akabane, T., and Hamaguchi, S., "Robust word spotting in adverse car environments," *Proceedings EUROSPEECH '93*, vol. 2, 1045-1048, 1993.

[5] Brancaccio, A., and Pelaez, C., "Experiments on Noise Reduction Techniques with Robust Voice Detector in Car Environment," *Proceedings EUROSPEECH '93*, 1259-1262, 1993..

[6] Mokbel, C. and Chollet, G, "Automatic Word Recognition in Cars,", CNET Technical Report, 1993.

[7] Acero, A. and Stern. R.M.,"Environmental Robustness in Automatic Speech Recognition," *Proceedings ICASSP*, 49-852, 1990.

[8] Naik, J.M., "A Speaker Verification System for Caller Identity Verification in the Telephone Network," *Proc. ECSA-ETRW Workshop on Speaker Identification and Verification*, Martigny, Switzerland, April 5-7, 1994.