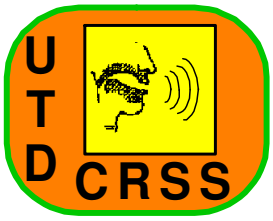
B. Pellom, J.H.L. Hansen, "Text-Directed Speech Enhancement using Phoneme Classification and Feature Map Constrained Vector Quantization," IEEE ICASSP-96: Inter. Conf. on Acoustics, Speech, and Signal Processing, vol. II, pp. 645-648, Atlanta, Georgia, May 1996.

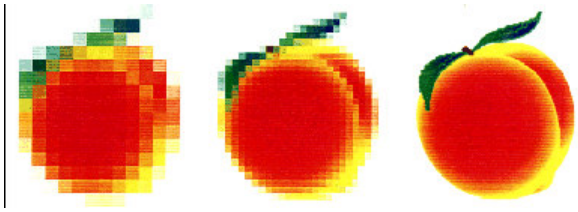
Text-Directed Speech Enhancement using Phoneme Classification and Feature Map Constrained Vector Quantization

Bryan Pellom, John H.L. Hansen



Center for Robust Speech Systems
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
The University of Texas at Dallas
P.O. Box 830688, EC33
Richardson, TX 75083-0688
972 – 883 – 2910 (Phone) 972 – 883 - 2710 (Fax)
John.Hansen@utdallas.edu (email)

IEEE ICASSP-96: Inter. Conf. On Acoustics, Speech, and Signal Processing, Atlanta, Georgia, May 1996.



The 1996 International Conference on Acoustics, Speech, and Signal Processing

TEXT-DIRECTED SPEECH ENHANCEMENT USING PHONEME CLASSIFICATION AND FEATURE MAP CONSTRAINED VECTOR QUANTIZATION*

Bryan L. Pellom and *John H.L. Hansen*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech bp@ee.duke.edu jhlh@ee.duke.edu

ABSTRACT

This paper presents and evaluates a novel text-directed speech enhancement algorithm for usage in non real-time applications. In our approach, the text of the intended dialogue is used to partition noisy speech into regions of broad phoneme classifications. Classes considered include stops, fricatives, affricates, nasals, vowels, semivowels, diphthongs and silence. These partitions are then used to direct a new vector quantizer based enhancement scheme in which class directed constraints are applied to improve speech quality. Objective enhancement evaluations conducted across 100 sentences of the TIMIT database indicate consistent improvement in speech quality for actual helicopter fly-by noise, aircraft cockpit noise, and automobile highway noise at signal-to-noise ratios ranging from -5 to 10 dB. Subjective quality assessment was conducted in the form of an A-B comparison test. Results of these evaluations demonstrate that, for wideband noise distortion, the proposed algorithm is preferred over unprocessed noisy speech more than 2 to 1, while the proposed algorithm is preferred over spectral subtraction processed speech by more than 3 to 1.

1. INTRODUCTION

Although most speech enhancement applications require real-time operation, many situations arise in which post-communication enhancement is vital. Example applications include the enhancement of black box recordings during aircraft crash investigations and the enhancement of 911 emergency calls for courtroom listening. In such instances, it is desired to improve quality as well as maintain intelligibility of prerecorded speech for which a text transcription (or some hypothesis driven text string) of the intended dialogue is available prior to enhancement.

Traditional speech enhancement algorithms such as spectral subtraction [2] and iterative Wiener filtering [7] fail to fully utilize the phonetic information carried within the signal for enhancement processing. This is a weakness since it is well known that noise has a non-uniform impact across the phoneme sequence of a speech utterance. As a consequence, the performance of these algorithms is limited primarily as a result of artifacts which are often introduced into the enhanced waveform. In fact, in extremely high noise environments, these artifacts may become just as overwhelming as the original degradation itself and further lead to losses in intelligibility.

Recent studies have demonstrated that improved enhancement can result when the phonetic inventory of the signal is considered during enhancement processing. One such study included the adaptation of the terminating iteration of the constrained iterative enhancement algorithm proposed by Hansen and Clements [5, 6]. In this scheme, a noise trained Markov model based phoneme classifier was used to partition speech into broad classes which in turn were used to adjust the terminating iteration of the algorithm on a frame by frame basis. Still other more perceptually motivated algorithms have been used to demonstrate improved enhancement performance when the speech signal was assumed to be comprised of voiced, transitional, and unvoiced regions [4].

In this paper, we address the problem of enhancing prerecorded speech for which it is assumed that both the text of the intended dialogue is known and the gender of the speaker is identified prior to enhancement. The paper is outlined as follows. Section 2 describes the formulation of the phoneme class partitioning scheme as well its integration into the enhancement algorithm. Section 3 details results of algorithm evaluations. This includes an evaluation of the broad phoneme class partitioning scheme as well as objective and subjective quality assessment of the text-directed enhancement algorithm. Finally, Section 4 summarizes findings and presents conclusions.

2. ALGORITHM FORMULATION

2.1 Text-Directed Partitioning

The text-directed partitioning algorithm is formulated as follows. The original 61 phonetic units from TIMIT were partitioned into 8 broad phoneme classes consisting of stops, fricatives, affricates, nasals, semivowels, vowels, diphthongs, and silence. 700 TIMIT sentences were downsampled to 8kHz and used to train hidden Markov models (HMMs) for each broad phoneme class. A five state, no-skip, HMM topology was chosen to model each of the eight phoneme classes. For each model, 2 Gaussian mixtures per state were employed. Speech training data was parameterized using the first 8 Mel frequency cepstrum coefficients, 8 delta Mel cepstrum coefficients, and frame energy. Model training was

*This work sponsored in part by the National Science Foundation and The Whitaker Foundation.

0-7803-3192-3/96 $5.00©1996 IEEE

accomplished using six iterations of the Baum-Welch algorithm.

To find the phoneme-class boundaries within a speech utterance, the text of the dialogue is first converted into a sequence of phoneme class labels. This is accomplished by using a dictionary of phonetic pronunciations. From this phoneme class sequence, a composite HMM is constructed by concatenating unit HMMs for each phoneme class in the sequence. During the alignment process, the input speech is first enhanced to attenuate the background noise prior to feature extraction. The front-end enhancement is accomplished using a constrained iterative technique known as (Auto:I,LSP:T) [6]. The enhanced speech is then parameterized on a frame by frame basis into a series of observation vectors. Using the Viterbi algorithm, the maximum likelihood state sequence is determined, and phoneme class boundaries are marked for each transition from the final state of the proceeding unit HMM model to that of the succeeding model along the maximum likelihood state path.

2.2 Enhancement Formulation

The enhancement algorithm consists of a non-iterative Wiener filtering scheme in which the power spectrum of the clean speech signal is estimated from a gender dependent vector quantizer codebook of clean LPC speech vectors. To improve speech quality, we utilize a robust distance metric, constrain the vector quantizer (VQ) search space based upon broad phoneme classifications, and apply inter-class continuity constraints to the estimated speech power spectrum across time in order to reduce residual artifacts in the enhanced speech waveform.

2.2.1 Class Constrained VQ-codebook Formulation

The 700 sentence TIMIT training corpus previously described was parameterized using the first 10 LPC coefficients chosen from windows of width 30 msec and an overlap of 75% (7.5 msec frame rate). The resulting clean speech vectors were partitioned by gender and quantized to form two 128 element VQ-codebooks (one male, one female).

The distance between test vectors and codebook entries was defined using the cepstral projection measure developed by Mansour and Juang [8]. This robust distance metric exploits the property that noise corrupted cepstral vectors are less sensitive to angle perturbation. The cepstral projection measure is defined as,

$$d(\vec{C}_r, \vec{C}_t) = |\vec{C}_t| - \frac{\vec{C}_t^T \vec{C}_r}{|\vec{C}_r|}, \quad (1)$$

where $\vec{C}_r$ represents the (clean) reference vector and $\vec{C}_t$ represents the (noisy) test vector. To facilitate the usage of the cepstral projection measure, each VQ-codebook LPC vector is mapped to a corresponding precomputed LPC cepstral representation. However, since lower order cepstral terms are more susceptible to noise than higher order cepstral terms, a liftering function as described in [8] is used,

$$w(i) = \begin{cases} 1 + 6.5 \sin\left(\frac{\pi i}{16}\right) & 2 \le i \le 12 \\ 0 & i = 1 \end{cases}$$

In the presence of high levels of background noise, choosing appropriate codebook entries which match the spectral characteristics of the intended clean speech can become quite difficult. Even more problematic are instances in which codebook entries selected to best represent the estimated clean speech spectra contain mismatch in formant location or bandwidth. Errors such as these can lead to perceptually annoying artifacts and further degrade the quality of the processed waveform.

In order to reduce such errors, the search space of the gender dependent VQ-codebook is constrained on a frame by frame basis by using the classifications provided by the parsing scheme. For example, frames classified as nasals are compared only to codebook entries found most likely to represent nasal sounds. The search space and constraint level are defined as follows. Let S_j represent the set of all codewords contained in the VQ search space for phoneme class j, given VQ-codebook C (i.e., $S_j \subseteq C$). Let $\vec{y}$ represent a test vector from a set of training data with prelabeled phoneme classification $\vec{y}_{CLASS}$. Furthermore, let D represent a codebook distance metric such that the closest codeword $\vec{c}_{min}$ in C from $\vec{y}$ can be written as,

$$\vec{c}_{min} = \arg\min_i (D(\vec{y}, \vec{c}_i)) \quad \text{for} \quad \vec{c}_i \in C. \quad (2)$$

Then a phone class dependent search constraint parameter ζ_j can be defined such that,

$$\zeta_j = \hat{p}\,(\vec{c}_{min} \in S_j | C, \vec{y}, \vec{y}_{CLASS} = j)\,, \quad (3)$$

with the condition that the number of codewords in S_j is minimized for fixed ζ_j. Here, ζ_j is fixed at run-time for each broad speech classification.

Essentially, ζ_j represents the estimated likelihood that the closest codebook entry from an arbitrary test vector $\vec{y}$ of class j will be a member of the constrained VQ search space, S_j. Values of ζ_j near zero indicate tight constraints, or a small search space, while values of ζ_j near one indicate relaxed constraints, or a large search space. The condition on S_j for a particular value of ζ_j indicates the method by which codebook entries are selected to comprise the search space. For example, with the condition that the number of codewords in S_j be minimized for a fixed ζ_j, the members of the search space are determined based upon selecting a series of highest probable codewords.

2.2.2 Speech Spectrum Estimation Technique

After the VQ-codebook search space has been constrained, the remaining codebook entries are compared with the current test vector using the cepstral projection measure. The *NBest* closest codebook entries are weighted to determine the estimate of the speech power spectrum as follows,

$$P_s^{(i)}(\omega) = \frac{g^2}{\sum_{n=1}^{NBest} w(n) \left|1 - \sum_{k=1}^{10} a_{n,k} e^{-j\omega k}\right|^2}, \quad (4)$$

where g represents the gain of the all-pole speech model, $a_{n,k}$ represents the *kth* LPC coefficient from the *nth* closest codebook vector, and $w(n)$ is an exponentially decaying weighting factor.

In order to reduce artifacts and impose a level of naturalness to the rate at which the speech spectrum is allowed to change across time, we obtain the speech power spectrum estimate at frame i, $\hat{P}_s^{(i)}(\omega)$, as a linear combination of spectra from the current frame and that of the previous frame. Here, the weighting factor ϕ is a constant dependent upon the phone class of the labeled frame.

$$\hat{P}_s^{(i)}(\omega) = \phi P_s^{(i)}(\omega) + (1-\phi)\hat{P}_s^{(i-1)}(\omega) \quad \text{for} \quad 0 < \phi < 1$$

At phone class boundaries, the constraint is not applied to allow for sharper transitions between classes. In this manner, transitions between stops and vowels, for example, are not overly smoothed. Finally, the *ith* degraded speech frame is filtered using the following Wiener filter,

$$H_i(\omega) = \left(\frac{\hat{P}_s^{(i)}(\omega)}{\hat{P}_s^{(i)}(\omega) + \alpha_i P_n(\omega)} \right)^{\frac{1}{2}},$$

where $P_n(\omega)$ represents the estimated noise power spectrum which is obtained during the silence frames, and the attenuation factor α, is adjusted on a frame-by-frame basis using the relationship,

$$\alpha_i = \arg\max \left(1, \frac{\alpha_{max}}{1 + SNR_i}\right),$$

where SNR_i is the estimated signal-noise-ratio for the *ith* degraded frame, and α_{max} was set to 30 for all evaluations presented in this paper. This attenuation rule attempts to attenuate noise for stops and fricatives which may be more corrupted than high-energy speech classes such as vowels and diphthongs.

3 ALGORITHM EVALUATION

The parsing algorithm and objective quality evaluations were conducted using 100 sentences (72 male, 28 female speakers) taken from the TIMIT database. Four additive noise sources were considered. This included flat channel noise (FLN), aircraft cockpit noise (AIR), automobile highway noise (HWY), and helicopter fly-by noise (HEL).

3.1 Classifier Evaluation

The text-directed speech partitioning algorithm was evaluated as follows. For each speech waveform of the evaluation corpus, the hand-labeled 8kHz sampled phonetic transcription from the TIMIT database was converted into one of eight broad phoneme classes. Furthermore, the time-alignments of the transcription were removed in order to generate a phoneme label sequence. Using this sequence of phoneme classes, a composite HMM network was constructed. Prior to the alignment procedure, each speech waveform was corrupted with one of four additive noise sources (i.e., AIR, HEL, FLN, HWY) at global signal-to-noise ratios of 10,5, and 0dB. As described in Section 2, the original degraded waveform is enhanced prior to alignment using the (Auto:I, LSP:T) constrained iterative enhancement algorithm. For each aligned utterance, the estimated phoneme class transcription was compared against the hand-labeled TIMIT boundaries from which the mean misalignments (in milliseconds) were recorded for each noise level and each noise source. Results of parser evaluations are presented in Table 1. Here we note that the (Auto:I,LSP:T) algorithm improves partitioning alignment in the presence of noise by reducing the overall mean misalignment.

3.2 Objective Enhancement Evaluation

Objective quality assessment of the proposed algorithm was conducted as follows. For each noise source (i.e., AIR, HEL, FLN, HWY), global signal-to-noise ratios of 10, 5, 0, and -5 dB were examined. Objective quality evaluations were conducted using the Itakura-Saito (IS) likelihood measure

Mean Misalignment vs Global SNR (in msec)				
Noise	*Front-End*	*10dB*	*5dB*	*0dB*
FLN	None	78	153	282
	(Auto:I,LSP:T)	36	46	77
HEL	None	63	96	156
	(Auto:I,LSP:T)	45	70	94
AIR	None	58	80	144
	(Auto:I,LSP:T)	36	42	63
HWY	None	58	74	80
	(Auto:I,LSP:T)	35	46	60

Table 1: Comparison of phoneme class parser boundary detection under four actual additive noise sources with and without (Auto:I,LSP:T) front-end processing.

and changes in objective quality were examined across broad phoneme classifications. Table 2 illustrates quality improvement for speech originally corrupted by aircraft cockpit noise at 0dB SNR. Substantial overall quality improvement in voiced and unvoiced sections of speech is demonstrated for each noise source in Figure 1. This improvement is demonstrated by a reduced mean distortion value in IS measures compared to the original degraded speech. For purposes of comparison, evaluations with traditional spectral subtraction [2] are also included.

Aircraft Cockpit Noise (0 dB Global SNR)				
	Itakura-Saito Likelihood Measure			
Sound Type	*(a)*	*(b)*	*(c)*	*#frames*
Silence	6.08	6.77	0.93	5087
Vowel	2.78	0.83	0.52	13402
Nasal	3.23	3.92	1.80	2258
Stop	7.13	3.61	1.25	6584
Fricative	16.76	7.09	2.03	5367
Semivowel	4.49	2.00	1.17	3308
Voiced + Unvoiced	6.35	2.86	1.10	30919
Total	6.31	3.41	1.08	36006

Table 2: Objective quality assesment across broad phoneme classes for 100 sentences degraded by aircraft cockpit noise at 0dB SNR for (a) original degraded (b) enhanced using spectral subtraction, and (c) proposed algorithm.

3.3 Subjective Enhancement Evaluation

Subjective quality evaluations were conducted in the form of an A-B comparison test. In particular, subjects were presented with 2 sets of 15 sentence pairs. All sentences were taken from the TIMIT database and downsampled to 8kHz. In total, these 15 sentences were comprised of speech spoken by 7 female and 8 male adults. In the first set of 15 pairs, subjects were presented with noisy unprocessed speech degraded with flat communications channel noise at a global SNR of 5 dB as well as the same noisy speech after enhancement with the text-directed approach. For the second set of 15 sentence pairs, subjects decided preference for either speech enhanced with spectral subtraction [2] or the proposed text-directed enhancement algorithm.

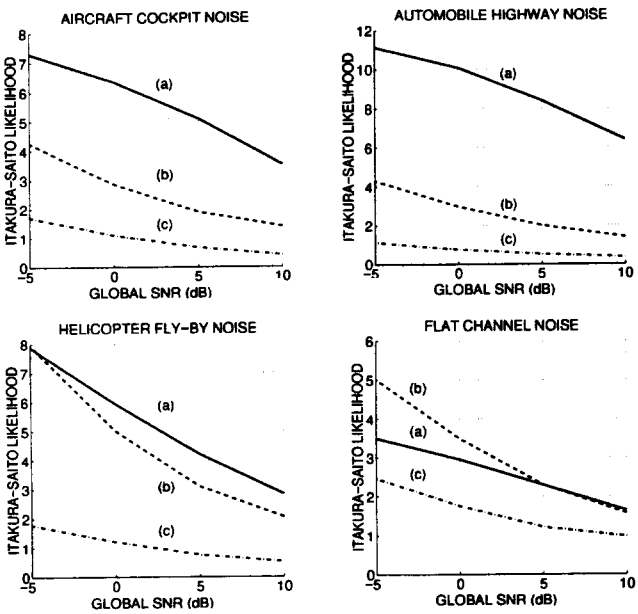


Figure 1: Mean Itakura-Saito Likelihood Measure across 100 sentences for (a) degraded, (b) spectral subtraction enhanced, and (c) proposed algorithm.

A total of 20 listeners took part in the quality assessment evaluation. Of these, 3 listeners were familiar with the speech enhancement field, while the remaining 17 were not familiar with any enhancement approach. All listeners possessed normal hearing, and claimed to have no history of hearing impairment. The results of the preference test were computed as follows. For each listener, the percentage of votes for one type of processing was compared to the percentage of votes for the remaining processing type. Furthermore, it was decided that a listener has a preference for processing method A over B if the listener prefers A more than 60% of the time. A listener is said to have no preference for either method if their preference lies between 40% and 60%.

Using the scoring definitions above, the results of the listening evaluations indicate that 12 listeners preferred the text-directed approach over the unprocessed noisy speech, 6 listeners preferred the noisy speech, and 2 listeners showed no preference. These results are encouraging in light of the objective quality evaluations presented in Figure 1. Here, we note that the objective quality improvement for the text-directed approach is not as substantial for wideband noise distortion, while a much greater improvement is demonstrated for the remaining noise sources. This would suggest that preference for the text-directed approach should be even greater for the remaining three noise sources.

When compared to linear spectral subtraction in the same setting, the text-directed approach favored well. In particular, 13 listeners preferred speech processed with the text-directed approach, 4 listeners preferred spectral subtraction, while 3 listeners showed no preference for either approach. However, one listener who preferred spectral subtraction over the text-directed approach felt that the musical tone artifacts in the spectral subtraction processed speech were "interesting" to listen to. Nevertheless, of the 13 listeners who chose the text-directed approach, many concluded that the speech processed by the text-directed method sounded much more natural compared to spectral subtraction.

6 DISCUSSION

In this paper, an off-line enhancement approach has been presented in which processing is adapted based upon the spoken text of the degraded utterance. In this approach, speech was partitioned into regions of eight broad phoneme classifications using an HMM based alignment strategy. It was illustrated that improvements to phoneme boundary alignments can be obtained via front-end enhancement processing. Furthermore, a novel vector quantizer enhancement scheme was developed in which phoneme classifications of each labeled frame were used to adapt the search space of the VQ-codebook. This was motivated in order to improve speech spectral estimation in the presence of noise, and reduce the potential for severe mismatch in spectral characteristics in the estimated spectrum. Additionally, a temporal continuity constraint was utilized to provide consistency across time in the enhanced waveform.

The algorithm was evaluated using both objective as well as subjective quality assessment techniques. Here, the text-directed approach was shown to improve the quality of degraded speech over a broad range of actual additive noise sources and signal-to-noise ratios. In each case, the proposed method was shown to lead to improved objective quality over linear spectral subtraction. Subjective quality assessment was conducted in the form of an A-B comparison test. Here, results show that, for wideband noise distortions, the proposed algorithm was preferred over the unprocessed noisy speech more than 2 to 1, while the proposed algorithm was preferred over spectral subtraction by more than 3 to 1.

References

[1] L. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptive Noise Suppression," *Proc. 1995 IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, pp. 812-815.

[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction.," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, April 1979.

[3] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Transactions on Audio and Electroacoustics*, 16(2):165-168, 1968.

[4] S. Nandkumar, J.H.L. Hansen, "Dual-Channel Iterative Speech Enhancement with Constraints Based on an Auditory Spectrum," *IEEE Transactions on Speech & Audio Processing*, vol. 3, no. 1, pp. 22-34, January 1995.

[5] J.H.L. Hansen, L. Arslan, "Markov Model Based Phoneme Class Partitioning for Improved Constrained Iterative Speech Enhancement," *IEEE Transactions on Speech & Audio Processing*, vol. 3, no. 1, pp. 98-104, January 1995.

[6] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795-805, April 1991.

[7] J.S. Lim, A.V. Oppenheim, "All-Pole Modeling of Degraded Speech," *IEEE Trans. on Acoust., Speech, Signal Processing*, pp.197-210, June 1978.

[8] D. Mansour, B. H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," in *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 11, pp. 1659-1671, November 1989.

[9] D. O'Shaughnessy, "Speech Enhancement using Vector Quantization and a Formant Distance Measure," *Proc. 1988 IEEE ICASSP*, pp. 549-552, New York, NY, May 1988.