

MINIMUM COST BASED PHONEME CLASS DETECTION FOR IMPROVED ITERATIVE SPEECH ENHANCEMENT

Levent M. Arslan and John H.L. Hansen

Digital Speech Processing Laboratory
Department of Electrical Engineering
Duke University, Box 90291
Durham, North Carolina 27708-0291

ABSTRACT

It is known that degrading acoustic noise influences speech quality across phoneme classes in a non-uniform manner. This results in variable quality performance for many speech enhancement algorithms in noisy environments. To address this, a hidden-Markov-model phoneme classification procedure is proposed which directs single channel speech enhancement across individual phoneme classes. The procedure performs broad phoneme class partitioning of noisy speech frames using a continuous-mixture hidden-Markov-model recognizer in conjunction with a cost based decision process. Cost functions are assigned which weigh errors between phoneme classes that are perceptually different (e.g., vowels versus fricatives, etc.). Once noisy speech frames are partitioned, iterative speech enhancement based on all-pole parameter estimation with inter and intra-frame spectral constraints (Auto:I,LSP:T) is employed. The phoneme class directed enhancement algorithm is evaluated using TIMIT speech data, and shown to result in substantial improvement in objective speech quality over a range of signal-to-noise ratios and individual phoneme classes. The algorithm is also shown to possess consistent quality improvement in a speaker independent scenario.

1. INTRODUCTION

Traditional enhancement methods employ the same basic processing approach throughout a sentence. However, since speech energy varies significantly across phoneme classes, local SNR as well as the impact of noise distortion on speech quality will vary locally. For example, vowel sections are not distorted to the same extent as transitive and plosive sounds for a given background acoustic noise level. Therefore, for speech enhancement to be successful, it must address the non-uniform effect noise has on speech across phoneme classes.

Improvements in speech enhancement could be achieved if additional *a priori* speech information is known or estimated prior to the enhancement process to address the

variable impact of noise. For example, Ephraim, Malah, and Juang [3] considered an approach where a hidden-Markov-model (HMM) recognizer is used to select an improved all-pole Wiener filter for enhancement. Another alternative, is to employ a vector quantization process as suggested by O'Shaughnessy [8], where a formant distance measure is used to select a noise-free entry from a vector quantizer codebook for enhancement.

It is suggested that such methods will result in increased *distortion* during enhancement, if an error is made in the decision process across time. In this paper, a cost based soft decision process is proposed for phoneme classification of noisy speech prior to enhancement. The enhancement approach considered is an extension of a previously formulated constrained iterative method by Hansen and Clements 1987,91 [4, 5].

2. ALGORITHM FORMULATION

The framework for the proposed enhancement algorithm is shown in Fig. 1. The technique is based on a speech class partitioning scheme which directs an (Auto:I,LSP:T) spectral constrained iterative enhancement algorithm. Noisy phoneme class partitioning is achieved using a continuous mixture HMM phoneme recognizer in conjunction with a decision process with cost functions that weigh errors between perceptually different speech classes. After class detection, each partitioned frame is enhanced using an appropriate set of constraints in the enhancement algorithm for each phoneme class. Each processing phase is discussed in the following sections.

2.1 (Auto:I,LSP:T) Enhancement

Consider a noise corrupted speech signal. It is assumed that input speech can be modeled by a set of all-pole parameters and a gain term. A basic speech enhancement technique is formed by performing a sequential maximum a posteriori (MAP) estimation of the clean speech given the noisy input speech, followed by MAP estimation of the model parameters given the speech signal resulting from the first MAP estimation [6]. This process iterates between estimation of the model parameters and estimation of speech signal, until a convergence threshold is reached.

¹This work sponsored in part by the Naval Research and Development, N66001-92-D-0092.

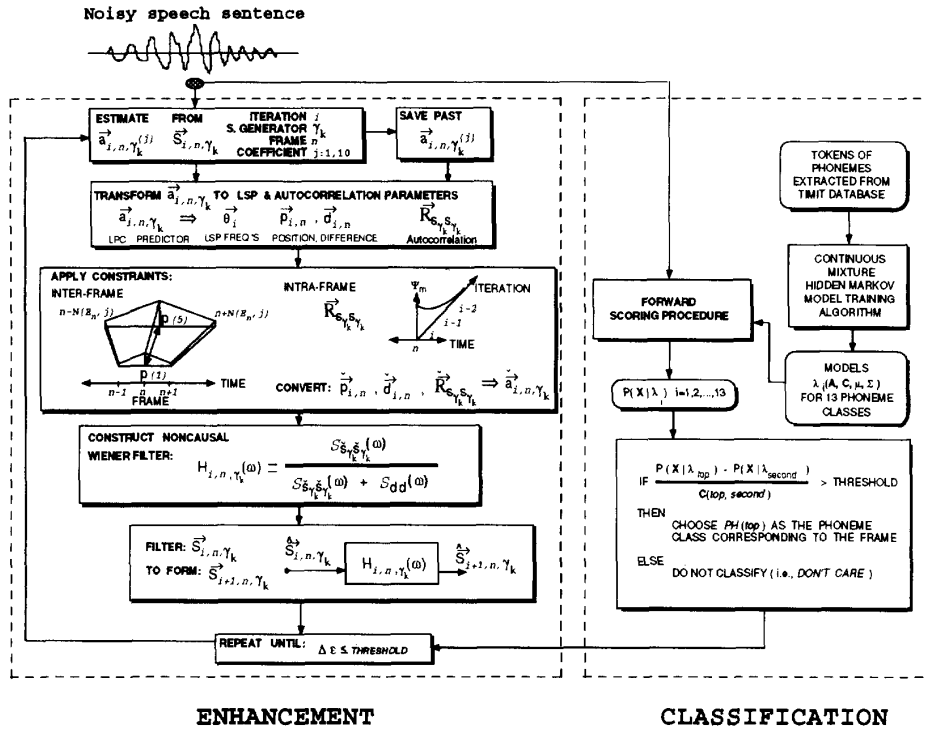


Figure 1: Framework for the classification directed (Auto:LSP:T) enhancement algorithm.

In order to improve parameter estimation, reduce frame to frame pole jitter across time, and provide a convenient terminating criterion, spectral constraints are introduced between MAP estimation steps on the line-spectral-pairs (LSP) and autocorrelation parameters, resulting in the inter- and intra-frame constrained (Auto:I, LSP:T) algorithm. (Fig. 1). This basic enhancement approach is discussed in further detail in [4, 5].

2.2 HMM Phoneme Class Partitioning

For classification, a training method based on continuous-mixture HMM is used to create 13 phoneme class models. The chosen classes² are: silence, closure stops, unvoiced stops, glottal stops, unvoiced fricatives, voiced fricatives, nasals, liquids, glides, diphthongs, front vowels, schwa vowels, and mid-vowels. The training algorithm creates models for each phoneme class using a 3 state left-to-right HMM with 5 mixtures weights. The models are trained using TIMIT [1] database sentences previously degraded with 10dB additive white Gaussian noise (AWGN).

Using phoneme class HMM models, the *forward algorithm* scoring procedure is employed to obtain the following conditional probabilities $P(\vec{X}|\lambda_i)$, $i = 1, 2, \dots, 13$ (i.e., the probability of phoneme class model λ_i producing the observation sequence \vec{X}). A phoneme class decision for each 64msec speech block is made using these probabilities grouped into seven broad classes (silence,

stops, fricatives, nasals, semivowels, diphthongs, vowels). The formulated decision process employs the two highest phoneme class probabilities. The difference between these probabilities is used to obtain a confidence measure for class decision. This is achieved by taking the ratio of $P_{top}(\vec{X}_i|\lambda) - P_{second}(\vec{X}_i|\lambda)$ to a cost value $C(top, second)$, where $C(top, second)$ is a measure of the cost of choosing $PH(top)$ when $PH(second)$ was the true class. If the confidence measure does not lie within a specified range, then that speech block is classified as a *DON'T CARE* (i.e., default general enhancement), which reduces the cost of a bad decision. Therefore, the HMM phoneme class partitioning is used to direct the enhancement algorithm only when the probability of a correct classification is high.

3 ALGORITHM EVALUATION

3.1 Classifier Evaluation

To determine performance of the classifier, TIMIT sentences degraded by AWGN were processed for classification. Using NIST [1] phoneme label data, correct and incorrect decisions were recorded. The gray scale image in Fig. 2 summarizes the resulting probabilities. The image, where black-to-white corresponds to $p : 0 \rightarrow 1$, shows the confusion matrix for class partitioning. The regions along the main diagonal represent correct decisions (e.g., classifying a nasal as nasal). Entries close to the main diagonal represent incorrect decisions which do not cause drastic degradation in the enhancement process (e.g., classifying

²The stated phoneme classes were used based on phonetic information files provided by NIST for the TIMIT database[1].

a semivowel as a diphthong). Entries far from the main diagonal correspond to poor, incorrect decisions which cause serious degradation in the subsequent enhancement procedure (e.g., classifying a fricative as a vowel). The figure shows that the classifier performs well since the concentration is high along or close to the main diagonal. In the evaluations, 30 percent of the frames were classified as *DON'T CARE*, resulting in the use of default (Auto:I, LSP:T) enhancement.

3.2 Enhancement Evaluation

An example of the algorithm's ability to adapt to the changing impact of noise on phoneme quality is illustrated in Fig. 3. Here, Itakura-Saito (IS) objective speech quality measures [7] are shown between original and noisy speech (10 dB SNR, AWGN) for the sentence "Often you'll get back more than you put in." The third plot is the corresponding distortion between degraded and original speech signals, which visualizes the non-uniform effect additive noise has on resulting speech quality. This is clearly illustrated for transitional and plosive sounds which are distorted more than vowels or even periods of silence.

The novel aspect here is that, the phoneme class directed speech enhancement algorithm attempts to reduce the peak distortion areas by applying a different terminating iteration according to the phoneme class under consideration. After a number of off-line simulations, the best terminating iteration for each phoneme class is determined. For *DON'T CARE* frames, the terminating iteration is set to the mean of the terminating iterations of all phoneme classes. The proposed algorithm performs constrained iterative speech enhancement (Auto:I,LSP:T) using phoneme class partitioning as a means to determine the best terminating iteration for each frame of speech data (henceforth referred to as (class partitioned(CP),Auto:I,LSP:T)). In Fig. 3, the fourth plot is the enhanced speech sentence using (CP,Auto:I,LSP:T), and the fifth presents the distortion measure between the enhanced and original speech signals. A comparison of the IS time based measure plots illustrate a dramatic reduction in noise distortion, and improved objective quality. Moreover, the resulting speech quality is seen to be primarily uniform over the whole sentence.

To confirm consistency of performance of the

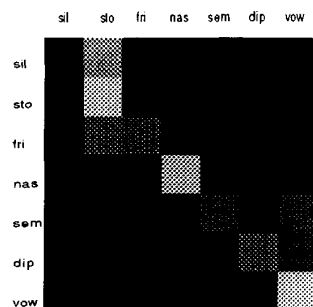


Figure 2: Confusion matrix for classification evaluation over a set of TIMIT sentences (Key: *silence, stop, fricative, nasal, semivowel, diphthong, vowel*).

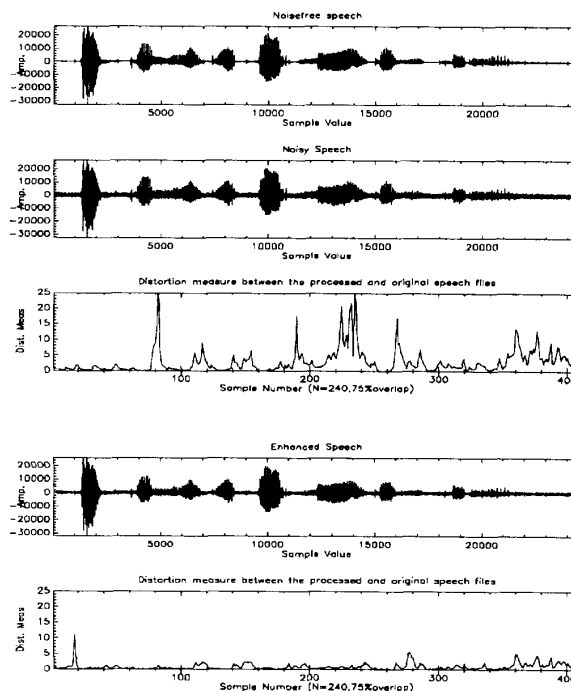
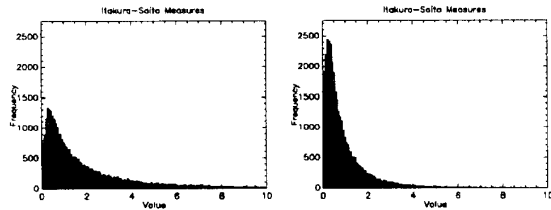


Figure 3: Time waveforms of the original, degraded, and enhanced speech sentence "Often you'll get back more than you put in." Distortion as measured by frame-to-frame objective quality measures are shown for noisy and enhanced waveforms.

(CP,Auto:I,LSP:T) algorithm, an evaluation was performed over 100 TIMIT sentences (72 male, 28 female) degraded with 10 dB SNR AWGN. Histograms of the IS distance measures corresponding to (a.) degraded, and (b.) enhanced for 100 sentences are shown in Fig. 4. The proposed enhancement technique reduced the mean distortion of the noisy speech considerably (i.e., from 2.750 to 1.003). The reduced tail of the enhancement quality histogram suggests that a majority of the frames result in a bounded level of distortion. This is encouraging given the fact that the (CP,Auto:I,LSP:T) method was tested in a noisy speaker independent environment. Informal listening tests also confirmed the increase in speech quality.

A comparison of the IS distance measures between degraded and enhanced speech for each broad phoneme class is illustrated in Fig. 5. In each plot the phonemes are sorted with respect to the level of quality improvement the algorithm introduced. The phonemes that resulted in the minimum and maximum improvement for each phoneme class are labeled with their TIMIT database representations [1]. There is considerable reduction in distortion as represented by the IS measures. The only exception was for fricatives, where the limited improvement can be explained by the fact that they are not initially distorted as much as other phoneme classes. The important point to note is the uniformity of the measures over all phonemes after enhancement. This fact suggests the usefulness of the proposed enhancement algorithm as a front-end to



(a.) Ave. IS dist. = 2.750 (b.) Ave. IS dist. = 1.003

Figure 4: The histogram of IS distance measures for 100 TIMIT sentences for (a.) degraded and (b.) enhanced using (CP,Auto:I,LSP:T) for 4 iterations.

Statistics of IS Distances over Phoneme Classes				
	DEGRADED		ENHANCED	
	Mean	Variance	Mean	Variance
Diphthongs	1.657	0.265	0.623	0.066
Nasals	7.600	1.025	1.513	0.039
Vowels	3.059	3.628	0.790	0.200
Stops	2.417	0.888	1.096	0.103
Fricatives	1.124	0.638	0.928	0.035
Silence	1.903	0.568	0.993	0.066
Semivowels	5.012	4.463	1.800	0.153
Overall	2.750	4.281	1.003	0.204

Table 1: Statistics of IS distance measures for previous 100 TIMIT sentences over broad phoneme classes.

other speech processing systems for recognition or coding in noisy environments. In Table 1, the means and variances of the distance measures are shown for each phoneme class. All phoneme classes were enhanced with the overall variance of distortion dropping from 4.281 to 0.204. This result is a strong indication of the success of the proposed algorithm, since the main objective was to eliminate the adverse effects of the nonuniformity of distortion over different phoneme classes by adapting the enhancement approach.

4. CONCLUSIONS

A phoneme class based partitioning process is formulated which directs the terminating iteration for an inter and intra-frame constrained (Auto:I,LSP:T) iterative enhancement algorithm. The algorithm addresses the non-uniform influence degrading acoustic noise has on speech quality across phoneme classes. A continuous-mixture hidden-Markov-model phoneme classification process, with cost based decision features is proposed which directs single channel speech enhancement across individual phoneme classes. Cost functions are assigned which weigh errors between phoneme classes that are perceptually different (e.g., vowels versus fricatives, etc.). The phoneme class directed enhancement algorithm is evaluated using TIMIT speech data and shown to result in substantial and consistent improvement in objective speech quality over a range of signal-to-noise ratios and individual phoneme classes.

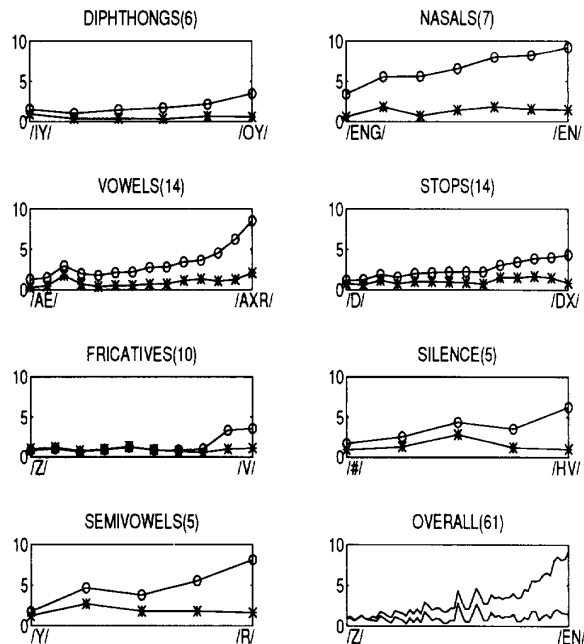


Figure 5: IS distance measures over broad phoneme classes for 100 TIMIT sentences for degraded (o), and enhanced (*) using (CP,Auto:I,LSP:T) for 4 iterations.

References

- [1] "Getting Started With The DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, (prototype as of December 1988).
- [2] L.M. Arslan, "Markov Model Based Phoneme Class Partitioning for Improved Iterative Speech Enhancement," M.S. Thesis, Dept. of Electrical Engineering, Digital Speech Processing Laboratory Tech. Report DSPL-93-6, Duke University, Durham, North Carolina, April 1993.
- [3] Y. Ephraim, D. Malah, B.H. Juang, "Speech Enhancement Based Upon Hidden Markov Modeling," *Proc. 1989 IEEE ICASSP*, pp. 353-356, Glasgow, Scotland, May 1989.
- [4] J.H.L. Hansen, M.A. Clements, "Iterative Speech Enhancement With Spectral Constraints," *Proc. 1987 IEEE ICASSP*, pp.189-192, Dallas, Texas, April 1987.
- [5] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. on Signal Proc.*, pp. 795-805, April 1991.
- [6] J.S. Lim, A.V. Oppenheim, "All-Pole Modeling of Degraded Speech," *IEEE Trans. on ASSP*, pp.197-210, June 1978.
- [7] S.R. Quackenbush, T.P. Barnwell, M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
- [8] 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.