ENHANCEMENT OF ESOPHAGEAL SPEECH BY INJECTION NOISE REJECTION

Hector Javkin, Michael Galler, and Nancy Niedzielski

Panasonic Technologies Inc., Speech Technology Laboratory
3888 State St., Santa Barbara, CA 93105, USA

ABSTRACT

Esophageal speakers, who produce a voice source by bringing about a vibration of the esophageal superior sphincter, must insufflate the esophagus with an air injection gesture before every utterance, thus creating an air reservoir to drive the vibration. The resulting noise is generally undesired by the speakers. This paper describes a method for the automatic recognition and rejection of the injection noise which occurs in esophageal speech.

1. INTRODUCTION

Persons who have had laryngectomies have several options for the restoration of speech, none completely satisfactory. The artificial larynx, typically a hand-held device which introduces a source vibration into the vocal tract by vibrating the external walls, is the easiest for patients to master, but does not produce airflow, so that the intelligibility of consonants is diminished. Tracheo-esophageal speech, which utilizes a prosthesis to divert outgoing lung air into the esophagus, bringing about a vibration of the esophageal superior sphincter, provides airflow for consonants and permits utterances of normal duration. However, it requires a surgically produced connection between the esophagus and the trachea, and is not suitable for some patients. Esophageal speech, which requires speakers to insufflate, or inject air into the esophagus[1], limits the possible duration between air injection gestures, and is associated with an undesired audible injection noise, sometimes referred to as an "injection gulp". The effect of this noise is magnified because esophageal speakers (like tracheo-esophageal speakers) evidence low vocal intensity [2] and frequently need amplification. This noise is undesirable for two reasons: (1) listeners and speakers find it objectionable and (2) in some speakers it can be mistaken for a speech segment, diminishing intelligibility. This paper reports on work to detect the injection noise, with the aim of eliminating amplification during its production.

To the best of our knowledge, this problem has never been addressed successfully, although considerable work has been undertaken to enhance other aspects of esophageal speech, particularly by Qi, Weinberg, Bi [3] [4], and colleagues.

2. OBJECTIVES

Since air injection is required prior to the start of every utterance and typically occurs after every pause before an utterance continues, it is possible to switch amplification on only after injection noise has occurred, and switch amplification off after a period of silence has occurred, while speech is transmitted without interruption. A gain control is set to either one or zero depending on whether injection noise has been detected with an associated silence, resulting in a device which is designed to automatically remove undesired injection noise.

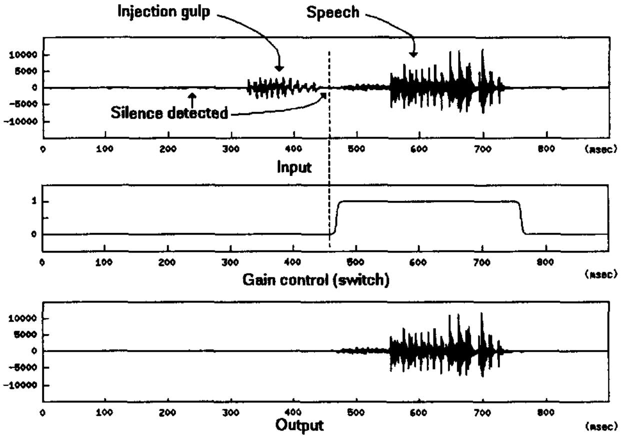

Figure 1. Characteristics of method for rejecting injection noise.

3. DETECTION OF NOISE BY HMMS

3.1. Feature analysis.

One method for detecting injection noise is based on relatively straightforward speech recognition and word-spotting methods. It essentially treats the injection noise as a word to be spotted. The basic scheme is shown in Figure 2.

The signal is digitized by sampling at 20 kHz. One copy of the signal is pre-emphasized and is used for processing, while a second copy is switched on or off depending on the analysis. Every 10 ms. a 256-point FFT computation is performed on a 20 ms. window of speech samples. The first 12 Mel-frequency cepstral coefficients (MFCC) are calculated; these form the first part of the feature vector for a speech frame.

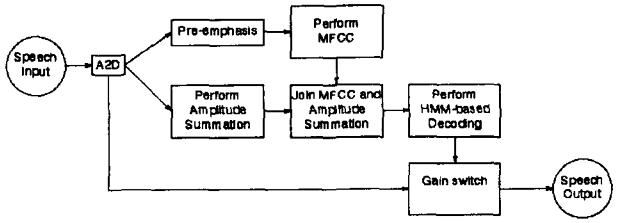
Speech Input
A2D
Pre-emphasis
Perform MFCC
Perform Amplitude Summation
Join MFCC and Amplitude Summation
Perform HMM-based Decoding
Gain switch
Speech Output

Figure 2. Flowchart of method for rejecting injection noise based on the recognition of silence and injection gulps.

This spectral information is supplemented by additional information about rate of change of spectral features, consisting of the derivatives (i.e. difference cepstra). All together, 24 Mel-based cepstral coefficients are extracted from each window of the speech signal.

Time waveform analysis supplements the cepstral analysis. Specifically, a measure of signal energy is computed, along with the energy rate-of-change, based on a linear regression of 9 successive samples.

The speech vector is further augmented with two extra feature points based on some special characteristics of the injection noise. When a voiced speech signal begins, it produces a negative pressure pulse. The injection noise, on the other hand, begins with a positive pressure pulse. The difference between the initial negative pressure pulse of speech and the initial positive pulse of the injection gulp is used to detect the injection gulp. A combination of a microphone, amplifiers and an analog converters is used to provide a non-inverted signal. This is done either by utilizing an even number of inverting amplifiers or by testing for an inverted signal and adding an inverting amplifier if necessary.

One of the features used to detect the polarity difference between injection noise and speech we have named Amplitude Summation (AS). Amplitude summation, computed once per 10 ms. speech window in the present implementation, is a means for detecting the initial deviation from zero of the speaker's signal. The digitized waveform is summed over intervals ranging from 1 to 20 milliseconds, depending on an adjustment for individual speakers. The probability that an injection gulp has occurred is greater when a positive value over a given threshold occurs in the summed signal. This threshold can be adjusted, depending on the associated microphones and amplifiers used to record the signal.

A second measure for detecting the polarity is differencing the center-clipped signal. To remove low-amplitude ambient noise, the signal is center clipped. The remaining signal is then differenced, to obtain the first derivative, which is then smoothed with a running average. A positive value on the result, immediately following a zero value, tends to indicate the presence of injection noise, while a negative value tends to indicate the presence of speech.

The three measures of signal energy, energy rate of change, and Amplitude Summation are added to the 24 Mel coefficients, to make up the complete observation vector. Thus, the acoustic front-end program creates a 27-component observation vector to represent the features of each speech frame.

3.2 Decoding

A Hidden Markov Model (HMM)-based speech decoder is used to find the optimal alignment of the speech signal with a set of speech tokens. Two methods are described.

In method one, five speech tokens are used, including silence, gulp, noise1, noise2 and speech. In method two, the speech token is replaced by a set of units representing the basic phonemes of the language. This method has more discriminative power for increased accuracy, but requires more computation.

Each token is modeled with an HMM. The number of nodes in the HMM units varies from 3, in the case of simple models such as silence, to as many as seven for certain phonemes. The number of gaussian densities per mixture may be varied from 6 to 18 or more, depending on the limits placed on computation time in the application.

In the first implementation, five continuous mixture density HMMs were trained on a subset of a corpus of esophageal speech data, segmented and pre-labeled by hand. The HMMs contained from 3 to 7 states, with 8 gaussian densities per mixture. The training procedure was initialized by training two models on an 8 kHz database of normal speakers: a speech model and a silence model. The distributions of these HMMs were then used to initialize the three other units. The five HMMs were then re-trained on the training half of the esophageal speech signals for the speaker, 42 recordings in all, using Baum-Welch re-estimation. This stage of speaker-adaptive training consisted of two iterations of isolated segment training and two iterations of non-segmented (i.e. embedded) training.

The HMM decoder program decodes the speech signal frame synchronously, with a 10 ms. advance rate. Each signal is processed by a front-end program into a vector of speech frames, as described in 3.1. The Viterbi algorithm [5] is used to estimate the probabilities of the speech token HMMs with respect to these feature vectors.

Finally, those segments for which the injection noise (gulp) token have been labeled as output are classified as gulps within the speech signal. The esophageal speech is transmitted with a short delay to permit processing, and amplified. When an injection gulp is detected, amplification is set to zero, so that they are not transmitted.

4. RESULTS OF DETECTION OF NOISE BY HMMS

The injection noise method was applied to a test set of 40 utterances on which the HMMs were not trained, but from the same speaker. The results are reported in Table 1.

Two thirds of injection noise, or gulp, events were detected successfully on the speaker. Of valid speech segments, 5.4% of them were at least partially incorrectly aligned with the gulp token (speech misclassification error). These results were obtained on 40 test sentences of one speaker. Although it is likely that these results can be improved by the use of more training data and further tuning of the recognition algorithm, some of the characteristics of the injection noise led to the exploration of a different approach.

Number of Speech Units	239
Number of Injection Gulps	72
Gulp-detection Error Rate	33.3% (24)
Speech Misclassification Error	5.4% (13)

Table 1. HMM Experimental Results

5. DETECTING ESOPHAGEAL INJECTION NOISE BY MORPHOLOGICAL FILTERING

A different method for injection noise detection has been developed, based on the observation that the noise, which is produced by a gesture with a closed vocal tract, has a strong, low-frequency emphasis. This characteristic appears to be due to a double closure in the vocal tract of at least some speakers, which strongly attenuates high frequencies.

It uses a simpler, faster, and more effective algorithm, which can be expected to become far more effective once it has been properly tuned. The data is sampled at 8 kHz. A 256-point FFT is computed, every 10 ms. and smoothed by a morphological filter [6, 7] with a 10 point sliding window, removing all but the gross features of the spectral curve.

Figure 3 shows the magnitude spectrum from the center of an injection noise segment and the output of the morphological filter.

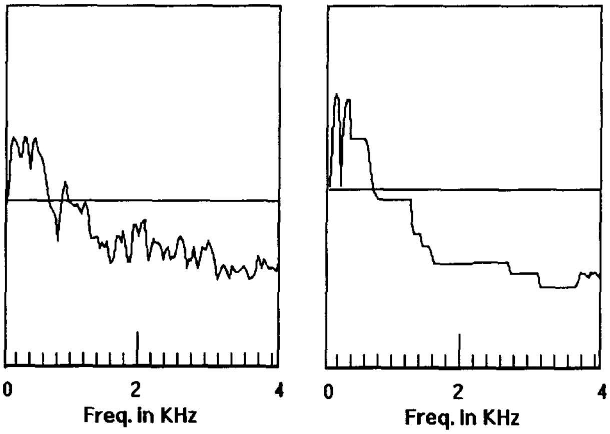

Figure 3. 256-point FFT from the center of an injection noise segment and the result of passing the FFT passed through the morphological filter.

Figure 4 shows the magnitude spectrum from the center of the consonant /d/ (the segment spectrally closest to an injection noise segment) and the output of the morphological filter (MF).

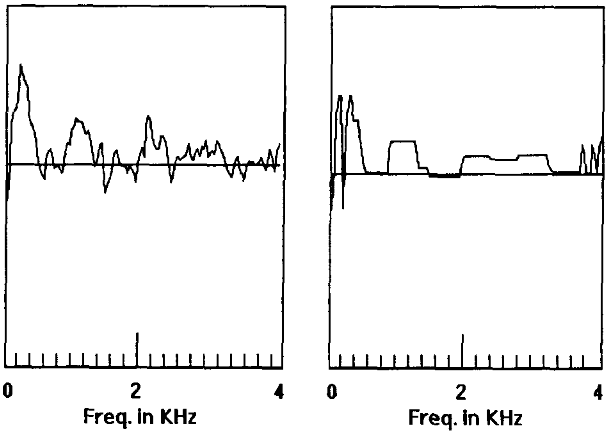

Figure 4. 256-point FFT from the center of a /d/ segment and the result of passing the FFT passed through the morphological filter.

The mean and the derivative of the filtered spectrum are computed. The location and value of the two largest peaks are identified. A signal segment is identified as injection noise if the following criteria are met:

a) The largest peak is lower in frequency than the second largest peak.

b) All points above 4 kHz are less than the mean.

The initial experiment testing this method was performed on both data sets used for training and testing the HMMs in the method described above. Again, results are promising, and point to the need for further development with more data.

Number of Speech Units	235
Number of Injection Gulps	79
Gulp Detection Error Rate	17.7% (14)
Speech Misclassification Error	15.7% (37)

Table 2. MF Experimental Results (Data Set 1)

Number of Speech Units	242
Number of Gulps	72
Gulp Recognition Error	16.7% (12)
Speech Misclassification Error	17.4% (42)

Table 3. MF Experimental Results (Data Test Set)

6. DISCUSSION

The results obtained thus far were obtained with relatively untuned algorithms. Furthermore, no attempt has been made at present to combine some of the features used in the HMM-based method with those used with the method based on morphological filtering. It is therefore likely that error reduction can be achieved. On the basis of these results and the likelihood that they can be improved, injection gulp rejection could work in a way totally transparent to the user, by means of an electronic switch that would only turn amplification on after a gulp and a following short silence have occurred. Whenever a speaker paused, the amplification would be turned off, waiting for the injection gulp before turning it on again. Although this method deals with only one aspect of esophageal speech, it could, in theory, work without any delay in the output signal.

Adjustments have to be made for speakers who use multiple gulps in order to sufficiently insufflate the esophagus. If a speaker consistently used double or triple gulps, the method could be tuned to reject them. However, speech with varying numbers of gulps could only have the initial gulp rejected.

REFERENCES

[1] Weinberg, B. & Bosna, J.F. Similarities between glossopharyngeal breathing and injection methods of air intake for esophageal speech. *Speech Hear Disord* 35:25-32, 1970.

[2] Robbins, J., Fisher, H.B., Blom, E.C., and Singer, M.I., A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Speech Hear Res* 49:202-210, 1984.

[3] Qi, Y., Replacing tracheoesophageal voicing sources using LPC synthesis, *J. Acoust. Soc. Am.* 88:1228-1235, 1990.

[4] Qi, Y., Weinberg, B. and Bi, N., Enhancement of female esophageal and tracheoesophageal speech, *J. Acoust. Soc. Am.* 98:2461-2465, 1995.

[5] Forney, G.D., The Viterbi algorithm, *Proceedings IEEE* 61, 268-278, 1973.

[6] Pitas, I. & Venetsanopoulos, A. N., *Nonlinear Digital Filters,* Kluwer Academic Publishers, Boston, 1990.

[7] Hansen, J. H. L., "Morphological Constrained Feature Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," *IEEE Trans. SAP,* vol. 2, pp. 598-614, 1994.