Isolated Word Speech Recognition Using Vector Quantization Techniques and Artificial Neural Networks

Jesus Savage, Carlos Rivera, Vanessa Aguilar

Facultad de Ingeniería
Departamento de Ingeniería en Computación
University of México, UNAM.
México City C.P. 04510
México

Abstract

In this paper we show how to combine speech recognition techniques based on Vector Quantization (VQ) together with Artificial Neural Networks (ANN). Speech recognition based only on vector quantization has proved its usefulness for isolated words speech recognition for small vocabularies (100-200) and one speaker, due to its easy implementation and its fastest calculation. One way to improve the performance of this technique is to add an ANN block that will fix the errors of the VQ recognizer.

To demonstrate the utility of this approach we tested it obtaining a 100% recognition rate compared to 97% with the VQ recognition alone.

Introduction

One of the main objectives for using digital signal processing (DSP) techniques in speech recognition is to take advantage of the redundancy that exists in speech, and thus reduce the amount of data and operations necessary to process it. During these operations a good set of features are obtained, that will be used to implement the recognition.

Some speech coding techniques using VQ are based on linear prediction coding (LPC). The vector quantizer is a set of LPC vectors that represents the spectrum shape of the signals to be encoded. These LPC parameters are the inverse filter gain squared σ^2 and the linear predictive coefficients A_i, $i = 0, ..., M$, with A_0=1.

$$H(z) = \frac{\sigma}{\sum_{k=0}^{M} a_k z^{-k}}, \tag{1}$$

We use the Levinson and Durbin algorithm for the generation of the LPC parameters. They represent an autoregressive model and they are called code words. These parameters represent the all pole filter speech production model. The collection of code words is called a codebook [1].

The codebook is designed from a training sequence that is representative of the speech to be encoded by the system. For codebook creation the generalized Lloyd algorithm is used [2]. After the codebook is created, each frame S_j of the voice signal to be encoded is compared with each of the LPC stored vectors C_i, and the spectrum shape is coded by identifying the LPC vector C_b that best represents S_j according to some distortion measure d.

$$d(S_j, C_b) = min(d(S_j, C_i)), \tag{2}$$

i=1,...,Size-Codebook

Distortion Measures

The human ear perceives loudness in a semilogarithmic way; thus one way to compare two speech spectra is by defining

$$V(f, \hat{f}) = log f(w) - log \hat{f}(w). \tag{3}$$

A distortion measurement that can be used is the set of norms defined by

$$d(f, \hat{f})^p = \int_{-\pi}^{\pi} \frac{dw}{2\pi} |V(f, \hat{f})|^p. \tag{4}$$

This set of norms is related to how humans perceive sound differences.

Some of the distortion measures that we are using are the Itakura-Saito (d_{IS}) and gain normalized Itakura-Saito (d_{GN}) [4]. One way to solve the linear prediction problem in speech is to use maximum likelihood estimation techniques, this leads to the Itakura-Saito (d_{IS}) distortion measure:

$$d_{IS}(f, \hat{f}) = \int_{-\pi}^{\pi} \frac{dw}{2\pi} [e^{V(f,\hat{f})} - V(f, \hat{f}) - 1]. \tag{5}$$

For two power spectra $f(w)$ and $\hat{f}(w)$, the d_{IS} distortion between them is

$$d_{IS}(f, \hat{f}) = \int_{-\pi}^{\pi} \frac{dw}{2\pi} [\frac{f}{\hat{f}} - ln\frac{f}{\hat{f}} - 1]. \tag{6}$$

For power spectrum estimates f and $\hat{f}$ that have the autoregressive (LPC) form

$$f(w) = \frac{\sigma^2}{|A(z)|^2}, \tag{7}$$

where

$$A(z) = \sum_{k=0}^{M} a_k z^{-k}, \tag{8}$$

and z=e^{iwn}.

The gain normalized Itakura-Saito distortion measurement is used when we are just interested in the spectral shape and not in the gain.

The d_{GN} distortion is given by

$$d_{GN}(f,\hat{f}) = d_{is}(\frac{f}{\sigma^2},\frac{\hat{f}}{\hat{\sigma}^2}) = \frac{\alpha}{\sigma^2} - 1, \quad (9)$$

where

$$\alpha = r(0)r_\alpha(0) + 2\sum_{n=0}^{M} r(n)r_\alpha(n), \quad (10)$$

$$r_\alpha(n) = \sum_{i=0}^{M-n} a_i a_{i+n}, \quad (11)$$

and where the $r(n)$ are the time domain autocorrelations of $f(w)$.

One of the advantages of using these distortions is that we do not need to perform any Fourier transformation in order to compare spectra. We just need to compare the autocorrelation function of the input signal and compute the previous equations with the autocorrelation values of the LPC coefficients. This equation can be implemented easily using fixed point digital signal processing [4].

Word Speech Recognition Using Vector Quantizers

In speech coding by VQ, a single codebook is designed from a training sequence that is representative of all speech to be encoded by the system. For speech recognition using vector quantization one codebook is used for each word of the recognition vocabulary. Each codebook is created from a training sequence containing repetitions of one vocabulary word. For example, a codebook for the word "one" would be designed by running the vector quantizer design algorithm on a training sequence of several repetitions of the word "one."

In order to recognize a spoken word using VQ, it is passed through each vector quantizer of the recognition vocabulary and the quantizers give a set of distortions $d_1(t), d_2(t), ..., d_M(t)$. These distortions correspond to the distance of the vector quantizers that best fit the frame t of the voice signal, for each quantizer, 1,...,M. The decision on what word was spoken is made by comparing the global distortions $D_1, D_2, ..., D_M$ from each quantizer, and by selecting the word of the vocabulary whose quantizer gives the smallest global distortion during t=1,...,T (see Fig.1). Where

$$D_i = \sum_{t=1}^{T} d_i(t)$$

We used the Itakura-Saito gain normalized distortion measure. Then the word j is selected if its global distance is minimum,

$$D_j < D_i$$

$$i = 1...M, i \neq j.$$

Speech Recognition Using LPC Vector Quantizers and Artificial Neural Networks

In order to improve the performance of the speech recognition based on VQ we add an ANN block at the outputs of each of the VQs, that is we feed the ANN with the global distances of each of the VQ that represent each of the words (see fig. 2). We used a neural network with 3 layers, one input layer, one hidden layer and one output layer, of 10, 20 and 10 neurons in the same order.

For each repetition of the training set we obtained the global distances that each VQ gives and feed them to the ANN, setting the output of the training word to .99 and all the remaining outputs to .01. For the ANN training we used gradient descent momentum and adaptive backpropagation techniques.

The decision on what word was spoken is made by comparing the outputs of the ANN $O_1, O_2, ..., O_M$, and by selecting the word of the vocabulary whose output gives the biggest value. The word j is selected if its ANN output:

$$O_j > O_i, \quad (12)$$

$$i = 1...M, i \neq j.$$

Experiments and Results

For isolated speech recognition speaker dependent we performed two types of experiments: using only the VQ technique and the combination of VQ and ANNs. The recognition vocabulary was the numbers in Spanish from cero to nine.

For training we use 10 repetitions of each word, and for the testing set we had also 10 repetitions of each word. For the testing set we got a recognition rate for the VQ alone of 97.97%, and for the VQ and ANN together 100%.

As we can see from the results the system had an excellent performance for one speaker and a small recognition vocabulary. The next step in the research is to test this system with multispeakers and a bigger vocabulary.

Bibliography

[1] Burton D.K., Shore. Isolated Word Speech Recognition Using Multi Section Vector Quantization Code Books. Naval Research Laboratory. July 1984.

[2] Linde J. Buzo A. An Algorithm for Vector Quantization Design. IEEE Trans. on Communication, January 1980.

[3] Gray R., Buzo A. Distortion Measures for Speech Processing. IEEE Trans. Acoustic, Speech and Sig. Proc. August 1980.

[4] Savage J., Herrera A. Isolated Word Speech Recognition Using Hidden Markov Models and Multi Section Vector Quantization Code Books. Proceedings Communications Theory and Applications, Scotland, September 1991.

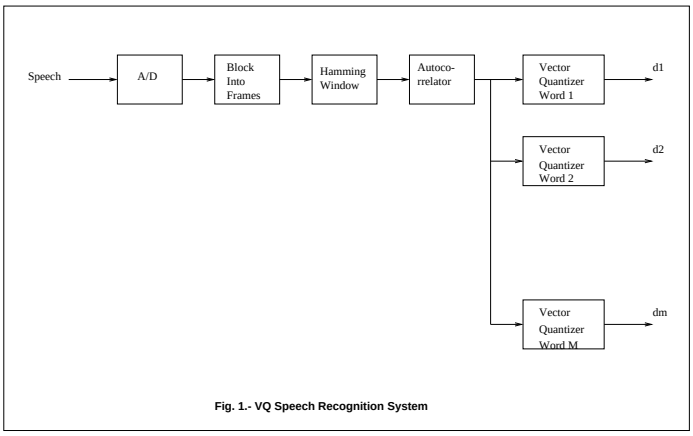

Fig. 1.- VQ Speech Recognition System

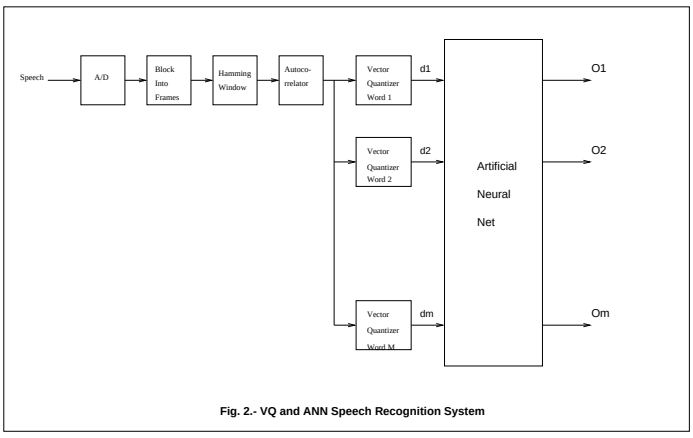

Fig. 2.- VQ and ANN Speech Recognition System