# NOISE ADAPTATION IN A HIDDEN MARKOV MODEL SPEECH RECOGNITION SYSTEM

Dirk Van Compernolle*

Department of Electrical Engineering - ESAT †

Katholieke Universiteit Leuven

Kardinaal Mercierlaan 94

B-3030 Heverlee

Belgium

Tel : (+32) 16/22.09.31

**Abstract**

Several ways for making the signal processing in an isolated word speech recognition system more robust against large variations in the background noise level are presented. Isolated word recognition systems are sensitive to accurate silence detection, and are easily overtrained on the specific noise circumstances of the training environment. Spectral subtraction provides good noise immunity in the cases where the noise level is lower or slightly higher in the testing environment than during training. Differences in residual noise energy after spectral subtraction between a clean training and noisy testing environment can still cause severe problems. The usability of spectral subtraction is largely increased if complemented with some extra noise immunity processing. This is achieved by the addition of artificial noise after spectral subtraction or by adaptively re-estimating the

*Research Associate of the National Fund for Scientific Research of Belgium (N.F.W.O.)

†The work was performed while the author was at IBM T.J. Watson Research Center, Yorktown Heights, NY 10598.

noise statistics during a training session. Both techniques are almost equally successful in dealing with the noise. Noise addition achieves the additional robustness that the system will never be allowed to learn about low amplitude events, that might not be observable in all environments; this, however, at a cost that some information is consistently thrown away in the most favorable noise situations.

# I. Introduction

The signal processing [1] used in IBM's real-time speech recognizer [2], has performed well in controlled environments. At the time of development the system was exclusively used in a quiet office with a typical signal to noise ratio of about 50dB. Day to day variations in a speakers voice level had to be accounted for, but the acoustic environment changed little.

The development of the portable Tangora system [9] allowed for use of the recognition system in a place at the convenience of the user, but simultaneously created a new set of problems. It became soon clear that the existing signal processing was not able to deal with the large variations in acoustic backgrounds that the system was now exposed to. Sensitivity to the absolute noise level was moderate, i.e. when the system was tested in the same acoustic environment as the one in which it was trained. Under those conditions a drop in signal to noise ratio from 50 to 20dB corresponded to a doubling in error rate. But a difference in the signal to noise ratio larger than 6dB, between the training and testing environment lead to severe deterioration in performance. Using a lip-mike can in principle solve most noise problems in office applications. The goal at IBM has been, however, to use the recognizer with a less constraining table mounted microphone.

A "silence model", representing the short pauses between words, is one of the hidden Markov models used in the isolated word recognition system. The inclusion of the silence model as a regular Hidden Markov Model (HMM) is necessary to avoid the otherwise difficult task of endpoint detection. It is trained in conjunction with the speech models and hence learns about the noise characteristics of the training session. Large differences in the noise level of training and testing environments will cause the recognizer to mistake speech for noise or noise for speech. Two rules for robust signal processing for speech recognition emerge : the processing of speech events must be largely invariant to the noise level and noise, independent of its actual value, should be mapped into a typical noise image. This is a normalization and adaptation task and therefore quite distinct from actual noise removal, which is the goal in speech enhancement [5,6].

The approach taken in this work was to maintain the general signal processing structure of the existing system, which has proven successful across many speakers in controlled constant acoustic environments, and to complement it with noise immunity processing that is largely transparent for the low noise environments but effective in dealing with changes in the background noise level. In section II the signal processing of the IBM speech recognition system is reviewed. In section III we introduce the basic noise immunity components of the signal processing which are spectral subtraction and a frequency dependent channel equalizer. Both operations rely heavily on a histogram based speech/noise discrimination algorithm. Refinements to the noise immunity signal processing and the statistical silence model are explored in sections IV and V. Ultimately in section VI results obtained on the IBM speech recognition system for all of the presented schemes are given and analyzed.

## II. Signal Processing Overview

The signal processing (Fig. 1) converts an input signal, sampled at 20kHz, into a 20 dimensional output vector at a frame rate of 100 frames/sec. It can be divided into 5 parts : Fourier transform, simulated filterbank, log conversion, long term adaptation and short term adaptation.

Figure 1: Signal processing block diagram (SIGPSTD)

All blocks, except for the long term adaptation, are identical to the design described in [1]. The FFT uses a 512 point Hanning window and creates a new spectral vector every 10msec. A mel-scaled filterbank is simulated by adding up FFT power spectrum coefficients. 20 channels spanning the frequency range from 200Hz to 7800Hz are created. The long term adaptation, which is described in detail in the next section plays a normalization role. Its output is ideally independent of the current acoustic environment. The short term adaptation is based on a Schroeder-Hall haircell model [7]. It is modeled after the rapid adaptation seen

in neural firing rate according to changes in input level. The time constants, dependent on the actual input, are between 30 and 50 msec. The importance of this block for the stationary parts of speech is minimal, but quite significant for the transient parts.

The 20 dimensional output vector is labeled by a vector quantizer with a codebook of size 200. These labels are the inputs to the HMM based recognition system [2]. The VQ codebook is designed by K-means clustering of training data [4]. This particular way of labeling, especially the use of a Euclidean distance metric in finding a closest prototype has a direct impact on the signal processing.

## III. Noise Correction and Acoustic Adaptation.

The function of the signal processing block "long term adaptation" is to compensate for the large day to day variations in the environment, and map the observed variable dynamic ranges into a fixed dynamic range about which the system is allowed to learn. The variations include changes in the background noise, room acoustics and recording hardware.

| | |
|---|---|
| $X^i$ | raw power spectral estimate |
| $\theta$ | rms noise threshold |
| $\mu^i$ | noise estimate |
| $SSTHR^i$ | spectral subtraction threshold |
| $G^i$ | channel gain |
| $y^i$ | adapted spectral estimate |

Figure 2: Basic long term adaptation for a single channel (NOISIM1)

The system proposed here (Fig. 2) [1] splits the correction for the noise and the adaptation to the room acoustics into two largely independent parts, but some linkage between both

[1] Notation : Capital letters are used for power spectrum variables, small letters for log spectrum variables. Channel numbers are given as superscripts.

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.