SPEECH VARIABILITY IN THE MODULATION SPECTRAL DOMAIN – SANOV A TECHNIQUE –

San r von Vuuren1,2
Hynek Herman s k y j,3

1 Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA
2 SpeechWorks International, Boston, Massachusetts, USA
3 International Computer Science Institute, Berkeley, California, USA

email: sar elv @ece.ogi.edu, hynek @ece.ogi.edu

ABSTRACT

This paper examines sources of variability in the speech signal using a new technique that is based on a nested spectral analysis of variance (SAN O V A). By constructing an ANOVA in the modulation spectral domain, the technique allows a characterization of unwanted variability in the time sequences of logarithmic energy caused by extraneous sources of variability such as additive noise, convolutional noise, and telephone handset transducer. Very low and moderate to high modulation frequencies are shown to be particularly affected by these sources. Verification results for 500 speakers on Switchboard data from the 1998 NIST speaker recognition evaluation are presented to confirm the conclusions. It is shown that a bandpass filtering and down sampling of the time sequences of logarithmic energy, compared to a conventional highpass filtering, leads to a 13% relative reduction of the EER in mismatched conditions.

1. INTRODUCTION

Many telephone based speech and speaker recognition systems use acoustic features based on filtered logarithmic energies derived from a short-term analysis [1, 2, 3]. This is done by estimating short-term logarithmic energies at given analysis frequencies om in the speech signal once every t_m = t_o - w_s = 5 or 10 ms, and deriving features by filtering the resulting time sequences of logarithmic energy. The effects and suitability of the filtering can be examined in the spectral domain, by using the power spectrum of each time sequence of logarithmic energy – a representation known as the modulation spectrum [4, 2, 5]. The filtering is motivated by the observation that a linear time-invariant communication channel affects the DC components in the time sequences of logarithmic energy. In particular, with a suitable choice of analysis window [2], the short-term logarithmic power spectrum of the transmitted speech signal S(n, o) = h(n) * s(n) can be decomposed in the n-th analysis segment as log|S(n, o)|2 ≈ log|H(o)|2 + log|S_e(n, o)|2, where H(o) is the frequency response of the channel and S_e(n, o) is the short-term frequency response of the speech signal before passing through the communication channel. Taking the Fourier Transform in the time variable n of the constant channel term log|H(o)|2, being independent of this time variable, is seen to affect the DC component of the modulation spectrum.

This paper extends the linear time-invariant framework to motivate a more general filtering. A technique is proposed that is based on a nested spectral analysis of variance (SANO V A) that allows a decomposition in the modulation spectral domain of various sources of variability in the speech signal. This allows variability to be analyzed across the range of components in the modulation spectrum and provides additional insight into the choice of features for speech and speaker recognition. Fig. 1 depicts the decomposition.

In telephone based speaker verification specifically, one type of variability that has been a major source of errors is caused by differences in handset type [6]. To illustrate, Fig. 2 shows two time sequences of logarithmic energy (each with mean removed) computed for one second of speech transmitted over both electret and carbon-button transducers.

To deal with the problem of handset variability, in [7] telephone handsets were classified according to electret- and carbon-button transducer type. One handset type was then mapped to the other using a memoryless polynomial nonlinearity so as to match short-term spectral magnitude. Among other things, the handset transducer was shown to create occasional "phantom-formants” in the spectrum of the speech signal. These were shown to occur at multiples and sums of the original formant frequencies and attributed to nonlinearities in the transducer.

By using the modulation spectrum of the time sequences of logarithmic energy we propose extending the analysis to
include medium-term\textsuperscript{1} effects of the handset transducers. Based on the preceding discussion, this is motivated by reasoning that the effects of handset variability will be time-varying, since the speech signal itself is time-varying.

The paper outline is as follows. The ANOVA technique is described in Section 2, Section 3 then presents a decomposition of telephone speech, while Section 4 shows how results from the decomposition can be used to improve speaker verification performance. Section 5 concludes.

2. TECHNIQUE

A nested analysis of variance (ANOVA)\textsuperscript{8} provides a convenient way to decompose the different sources of variability in the acoustic features. Defining $S(n, \omega_k)$ to be the short-term frequency response at the $k$-th analysis frequency, the decomposition is performed on the sequence of logarithmic energy $X(n, k) = \log(|S(n, \omega_k)|^2)$ and the contributions of the individual factors analyzed using the modulation spectrum. We next describe the data and analysis technique.

2.1. Data

The HTIMIT corpus which consists of speech from 192 males and 192 females is used. This corpus was collected by Douglas Reynolds\textsuperscript{9} and can be obtained through the LDC.\textsuperscript{2} It contains ten utterances per speaker, each of different spoken text. The HTIMIT utterances are the original TIMIT utterances of a speaker which were transmitted through different electret- and carbon-button transducers to simulate the effect of different telephone handsets. We used three different electret and three different carbon-button transducers and randomly selected 15 male speakers. Utterances were chosen so that the same text strings were spoken by each speaker (refer Fig. 1). We aligned the recordings for a particular speaker speaking the same text over different handsets using a waveform-based correlation and verified visually that alignment errors are on the order of 3 ms or less – i.e. less than a short-term analysis frame spacing of 5 to 10 ms.

2.2. Nested Spectral Analysis of Variance

Speech from different speakers speaking the same text are in general not finely aligned due to variations in rate of speech and prosody, as well as phoneme and syllable insertions, deletions and substitutions. This suggests that variations due to different speakers and text can not be interpreted separately as it would be difficult to disambiguate.\textsuperscript{3} However, this is not a problem for variations among handset transducers, since the speech signals associated with a particular speaker are aligned. The approach taken here is to estimate the handset and total variability with the understanding that the difference between the total and handset variability is the speech variability. The handset variability as a function of modulation frequency will be denoted as $G_H(\theta)$ and called the handset variation. The total variability as a function of modulation frequency will be denoted as $G_X(\theta)$ and called the total variation. The portion of the total variability that is not attributable to the handset variability will be denoted as $G_S(\theta)$ and called the speech variation.

In the analysis, all time sequences are truncated to the same length. The time sequence of logarithmic energy $X(n, k)$ with $1 \leq n \leq N$ is therefore treated as a $N$-dimensional vector. Without loss of generality, the frequency index $k$ is dropped from the notation. Except when mentioned otherwise, results are reported for the 1 kHz short-time analysis frequency. Nesting is on the factors handset transducer ($H$), speaker ($S$) and text ($T$) as shown in Fig. 1. It is assumed that the factors are i.i.d. and effects additive. The latter assumption is justified for a convolutional degradation as discussed in the introduction. The observed response at the $j$-th level of $T$, $l$-th level of $S$ and $i$-th level of $H$ is then

$$X(n) = \bar{X}(n) + X_j(n) + X_{H_l}(n) + X_{H_lT_i}(n) + \epsilon_{ijl},$$

with for $j = 1...J$, $l = 1...L(j)$, $i = 1...I(j, l)$ and where $\bar{X}(n)$ is the average response at time $n$. The experimental error is $\epsilon_{ijl}$ which we assume to be zero for the purposes of this study. The total variation can be obtained in terms of the sums of squares (SS) as

$$SS_{total} = SS_T(n) + SS_S(n) + SS_H(n).$$

Letting $L_j(l) = \sum_{i=1}^{T(j)} I(j, l)$, the individual terms are

$$SS_X(n) = \sum_{j=1}^{J} \sum_{l=1}^{L(j)} \sum_{i=1}^{T(j)} |X_{ijl}(n) - \bar{X}_j(n)|^2,$$

$$SS_H(n) = \sum_{j=1}^{J} \sum_{l=1}^{L(j)} \sum_{i=1}^{T(j)} |X_{ijl}(n) - \bar{X}_{ij}(n)|^2,$$

$$SS_S(n) = \sum_{j=1}^{J} \sum_{l=1}^{L(l)} |I(j, l) - \bar{X}_{ij}(n)|^2,$$

$$SS_T(n) = \sum_{j=1}^{J} \sum_{l=1}^{L(j)} L_j(l) |X_{ijl}(n) - \bar{X}_j(n)|^2.$$  \hfill (3)

The dot notation, as in $\bar{X}_{ij}$, is used to indicate the average value of $X$ computed over the "dotted" factors – in this example, $i$ and $l$. The speech variation $SS_{SP}(n)$ is derived from the speaker and text variations as

$$SS_{SP}(n) = SS_X(n) + SS_T(n) = SS_X(n) - SS_H(n).$$

To interpret the variations in the modulation spectral domain it is necessary to modify the computation. For example, the term $SS_X(n)$ for the total variation is modified as follows\textsuperscript{4}

$$G_X(\theta) = \sum_{j=1}^{J} \sum_{l=1}^{L(j)} \sum_{i=1}^{T(j)} |FT_n \{X_{ijl}(n) - \bar{X}_j(n)\}|^2,$$

where $FT_n$ denotes the Fourier Transform with respect to the time index $n$ and $\theta$ is the modulation frequency. The other terms $SS_S(n)$ and $SS_H(n)$ are modified in the same way.

\textsuperscript{1}From 10ms to a few seconds.

\textsuperscript{2}See LDC’s URL: http://www.ldc.upenn.edu

\textsuperscript{3}It may also not be possible to compensate for this ambiguity, since trying to do so may impose unnatural constraints on the speech signal and introduce artificial variability. This would happen for instance if a technique such as dynamic time-warping was applied to the signals.

\textsuperscript{4}Note that the modification preserves the overall energy as can be seen by applying Parseval’s Theorem

$$\sum_{n=\infty}^{\infty} |X[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X[\theta]|^2 d\theta$$

and by rearranging the summation.
way to obtain the variations $G_S(\theta)$ and $G_H(\theta)$. Without loss of generality $G_H$, $G_S$ and $G_X$ are normalized so that the total variation $G_X$ is approximately 1 at $\omega = 0$.

Eqn (5) can be seen to represent an analysis in the modulation spectral domain since it consists of suitable averages of the modulation spectra of components of the acoustic features. Recall that the modulation spectrum is the power spectrum of a time sequence of logarithmic energy. The equation requires estimates of the power spectrum which we compute using Welch’s averaged periodogram method using a 1 second long Hamming analysis window advanced in 1 second steps. The choice is based on the fact that the HTIMIT speech utterances, and therefore time sequences, are nominally 2.5 seconds in duration.

3. VARIABILITY

Fig. 3 depicts the total variation and handset variation on a logarithmic scale for time sequences obtained at a frame rate of $\theta_0 = 200$ Hz using a short-term Hamming analysis window of length $t_w = 20$ ms. The 1 second long Hamming analysis window that was used to estimate the modulation spectrum had the effect of smearing the energy in the DC component into components up to about 2 Hz. See [5] for a detailed discussion. The Nyquist frequency for the analysis is 100 Hz.

In the figure, the upper curve is the total variation $G_X(\theta)$ and the lower curve is the handset variation $G_H(\theta)$ as computed over the three electret- and three carbon-button transducers jointly. For modulation frequencies $\theta > 5$ Hz, it can be seen that the total variation falls off by about 7 dB per octave, whereas the handset variation falls off by about 3.5 dB per octave.

Fig. 4 shows the same result as a ‘signal-to-noise’ ratio

$$\frac{S}{N}(\theta) = \frac{G_S(\theta)}{G_H(\theta)} = \frac{G_X(\theta) - G_H(\theta)}{G_H(\theta)}.$$  \hspace{1cm} (6)

Results are shown for various filter bank bands with center frequencies $f$ spanning the telephone band. It can be seen that handset variability contributes considerably to the total variability at very low and moderate to high modulation frequencies ($\theta < 1$ Hz and $\theta > 10$ Hz). At the very low modulation frequencies this suggests a strongly varying DC component in the time sequences of the logarithmic energies, agreeing with a model for a convolutional degradation.

The technique can be used to analyze the effect of additive noise. For example, consider adding white noise at a fixed SNR to each electret speech signal.\textsuperscript{5} To ensure a fixed absolute noise level across different signals the broad-band energy of each speech signal is normalized prior to the analysis. Without loss of generality, this only affects the DC component in the logarithmic domain. Fig. 5 depicts the resultant total and handset variations for different SNR levels of additive noise. By comparing Figures 3 and 5 it can be seen that variations in broad-band energy contribute considerably to the DC variability. From Fig. 5 (a-c) it can been that adding noise to speech affects the SNR at all modulation frequencies and that increasing the additive noise leads to an increase in the handset variation at all frequencies with the effect that handset variation contributes substantially to the total variation at high modulation frequencies. This can be seen best in Fig. 5 (b-c) showing that the additive noise has a ‘pinching’ effect. It can be conjectured that attenuating the higher modulation frequency components may make the features less sensitive to such changes and perhaps increase system robustness.

4. SPEAKER VERIFICATION RESULTS

Speaker verification experiments were conducted to determine whether higher modulation frequency components are indeed affected deleteriously by variations among handsets. Features were obtained at a rate of 100 Hz as instantaneous and dynamic (5-point delta) filter bank outputs from which the long term averages have been subtracted. A baseline system used the features without modification, while the proposed system used features that were lowpass filtered to 10 Hz and down sampled to 25 Hz.

Results were computed over 30-second long speech segments of the 1998 NIST speaker recognition evaluation [6].

\textsuperscript{5}White noise from the NOISEX-92 corpus is used at an SNR measured in decibels by the ratio of average energy in the speech signal and average energy in the noise signal.
Training and testing handsets where chosen to differ in telephone number and transducer type (DNDT). Scores from 250 males and 250 female targets were pooled in computing a detection error tradeoff (DET) plot for the false acceptance and false rejection errors. As seen in Fig. 6, the proposed system performed better than the baseline system with a difference in performance that was statistically significant at the $p = 0.02$ level. The proposed system resulted on average in a relative reduction in equal error rate (EER) of 13%.

- Additive noise affects the modulation spectrum across the range of modulation frequencies. In particular, additive noise at SNR levels as low as 20 dB mask frequency components in the modulation spectrum of speech that are higher than about 20 Hz.

When applied to speaker verification the conclusions resulted in a relative reduction of EER in mismatched conditions.

**ACKNOWLEDGEMENT**

This work was supported in part by DoD grants MDA-904-97-1-0098 and MDA-904-98-1-0521.

**REFERENCES**


