Proc. of Workshop on Automatic Speech Recognition and Understanding, Keystone, Colorado, Dec. 1999.

TIME-FREQUENCY DISTRIBUTIONS FOR AUTOMATIC SPEECH RECOGNITION

Alexandros Potamianos * *and Petros Maragos* †

* Bell Laboratories, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07074, U.S.A.
† Dept. of ECE, National Technical University of Athens, Zografou 15773, Athens, Greece.

ABSTRACT

In this paper, the use of general time-frequency distributions as features for automatic speech recognition (ASR) is discussed in the context of hidden Markov classifiers. Specifically, short-time average of quadratic operators, e.g., energy spectrum, generalized first spectral moments, and short-time averages of the instantaneous frequency are compared to the standard front end features, and applied to ASR. Theoretical and experimental results indicate the equivalence of some of these feature sets and the close relationship among others. Future research directions in the area of feature extraction for ASR are suggested.

1. INTRODUCTION

Time-frequency distributions and short-time averages of quadratic operators are very popular front-end features for automatic speech recognition (ASR). Indeed, the "standard" front-end feature set is a short-time–frequency energy distribution. Despite the standardization of the ASR front-end, there has been a significant amount of research on using alternate time-frequency distributions as (possibly additional) ASR features. A good review of such efforts can be found in [9]. However, such efforts are often lacking in theoretical or experimental justification. In this paper we attempt to: (1) outline the relationships between some popular alternative feature sets and the "standard" front-end features, (2) present experimental ASR evidence that supports these claims, (3) outline possible fruitful research directions for ASR feature selection. We hope that this study will help guide future ASR front-end research.

The following two types of non-parametric features are investigated in this paper: (i) short-time averages of quadratic operators, e.g., energy spectrum [10], (ii) generalized first spectral moments and weighted short-time averages of the instantaneous frequency. Note that the standard feature set is included in the first family of time-frequency distributions. Our goal is to show (both theoretically and experimentally) a close relationship among these feature sets and the standard feature set.

The organization of the paper is as follows: First, we introduce the energy operator and the energy spectrum, and compare it to other spectral envelope representations. Then we propose short-time instantaneous frequency estimators in the context of the AM–FM modulation model, the sinusoidal model, and spectral estimation. The estimators are compared to the spectral envelope and their merits as ASR features are discussed. Finally, experimental ASR results are given and future research directions are proposed. The authors assume in the presentation some familiarity with the sinusoidal speech model, the AM–FM modulation model and energy operators.

Some of this work was performed while the authors were with the School of E.C.E, Georgia Institute of Technology, Atlanta, GA 30332, USA. It was partially supported by the US National Science Foundation under Grants MIP-9396301 and MIP-9421677.

2. QUADRATIC OPERATORS AND ENERGY SPECTRUM

The energy operator is defined for continuous-time signals $s(t)$ as

$$\Psi_c[x(t)] \triangleq [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \quad (1)$$

where $\dot{x} = dx/dt$. Its counterpart for discrete-time signals $s(n)$ is

$$\Psi_d[x(n)] \triangleq x^2(n) - x(n-1)x(n+1) \quad (2)$$

The nonlinear operators Ψ_c and Ψ_d were developed by Teager during his work on speech production modeling [13] and were first introduced systematically by Kaiser [2]. When Ψ_c is applied to signals produced by a simple harmonic oscillator, e.g. a mass-spring oscillator, it can track the oscillator's energy (per half unit mass), which is equal to the squared product of the oscillation amplitude and frequency; thus the term *energy operator*. The energy operator has been applied successfully to demodulation and has many attractive features such as simplicity, efficiency, and adaptability to instantaneous signal variations [3]. The attractive physical interpretation of the energy operator has led to its use as an ASR feature extractor in various forms, see for example [14, 15].

The *energy spectrum*, introduced in [10], is a general time-frequency distribution based on the energy operator. Assume that $x(n)$ is filtered by a bank of K bandpass filters centered at frequencies ω_k to obtain K band-passed signals: $x_k(n)$, $k = 1..K$. The following time and frequency relations hold

$$x_k(n) = x(n) * h_k(n) \leftrightarrow X_k(\omega) = X(\omega)H_k(\omega) \quad (3)$$

where $h_k(n)$ is the impulse response and $H_k(\omega)$ is the frequency response of the kth filter. The energy spectrum $ES(n,k)$ is defined as the short-time average of the energy operator applied to the family of band-passed signals $x_k(n)$, i.e,

$$ES(n,k) = \sum_{m=n}^{n+N-1} \Psi_d[x_k(m)] \quad (4)$$