Chapter 8.1
Literacy by Way of Automatic Speech Recognition

Russell Gluck
University of Wollongong, Australia

John Fulcher
University of Wollongong, Australia

ABSTRACT

The chapter commences with an overview of automatic speech recognition (ASR), which covers not only the de facto standard approach of hidden Markov models (HMMs), but also the tried-and-proven techniques of dynamic time warping and artificial neural networks (ANNs). The coverage then switches to Gluck's (2004) draw-talk-write (DTW) process, developed over the past two decades to assist non-text literate people become gradually literate over time through telling and/or drawing their own stories. DTW has proved especially effective with "illiterate" people from strong oral, storytelling traditions. The chapter concludes by relating attempts to date in automating the DTW process using ANN-based pattern recognition techniques on an Apple Macintosh G4™ platform.

INTRODUCTION: SPEECH PRODUCTION

Generally speaking, the aims of automatic speech recognition (ASR) are twofold: firstly, to extract the salient features of the incoming speech signals, then secondly to map these into the most likely word sequences, with the assistance of embedded acoustic and language models (Huang, Acero, & Hon, 2001).

Natural, conversational, continuous speech often incorporates false starts, repeated phrases, non-grammatical phrases (*ums* and *ahs*), and pauses, which bear little relation to written (text) punctuation. Some characteristics of speech which make recognition, whether by humans or machine, difficult include: background noise levels, variations in speaker loudness, pitch, emphasis (stress), and speech rate, not only *between* different speakers (either from within the same culture or due to different dialects), but also on different occasions with the *same* speaker (for

example, with or without a head cold). Even worse, we tend to make assumptions as to what words (phonemes) we *expect* to hear next, based not only on the context of surrounding words (phonemes), but also on cultural mores. Further, since there is not always a strong correlation between the acoustic properties of speech waveforms and the linguistic units that they represent, this can lead to ambiguous interpretation. Ambiguities can also arise due to the fact that similar-sounding words can have quite different meanings (homonyms); conversely, different-sounding words can have similar meanings (synonyms).

A person's fundamental frequency (number of vibrations per second) is a function of their vocal cord mass, and typically ranges between 50 and 250Hz for males, and roughly twice this frequency for females.

We generate speech (phones) using a combination of voice box, or larynx (the vibration source), lungs (energy or power source), vocal tract and nasal passage (resonant cavities), together with the articulatory organs (lips, teeth, tongue, jaws, cheeks, and alveolar ridge — that region in the roof of the mouth which makes contact with the tip of the tongue) (Masaki, 2000). The lips, teeth, tongue, jaw, and cheeks are all capable of changing the shape of the basic resonant cavity, thereby producing different sounds. For example, the lips are involved in the production of English vowels and the consonants /b/ and /p/; the teeth (and lips) in /f/ and /v/; the alveolar ridge in /d/, /n/ and /t/, and the cheeks in /b/ and /p/. Likewise, various constrictions in our air passageways produce different sounds (for example, /p/, /b/ and /f/). Furthermore, sounds can be produced either with the vocal cords vibrating, referred to as "phonation" or voiced (for instance, /g/, /m/, /z/), or without, in other words "voiceless" (such as /f, /k/, /p/, /s/, /t/) (Keller, 1994).

Thus from a signal processing point of view, we can regard speech as a time-varying sound wave, whose frequency components are determined by changes in the size and shape of the vocal tract and associated physiology. Peaks in the energy spectrum of the speech waveform are referred to as acoustic resonant frequencies or "formants". Most vowels comprise more than three formants; however, the first three (F1 ~500Hz, F2 ~1800Hz, F3 ~2500Hz), usually suffice for purposes of classification and/or recognition (higher-frequency formants reflect voice quality and individual speaker characteristics) (Ainsworth, 1997). Thus we can conceive of speech as the superposition of a number of frequency components of varying amplitudes and phases. As such, and in common with signal processing generally, speech is amenable to either Fourier series analysis (for continuous — analog — signals), or once digitized, to Fourier transforms (for discrete signals). Speech recognition is invariably implemented on some form of computer platform; thus the raw speech signal must first be converted from analog to digital form.

Acoustic signals, including speech, are characterized by features such as pitch, duration, amplitude (loudness, signal strength, power/energy), and phase of each frequency component. As it happens, only the first three are relevant from a speech recognition perspective, since the human ear is insensitive to phase (Katigiri, 2000). Now since phonemes, the basic linguistic unit, are characterized by frequency, time, and energy, it makes more sense to use three-dimensional spectrograms rather than process the raw (albeit filtered) time-varying speech waveform. Filtering is necessary since speech, like any other one-dimensional time-varying acoustic signal, is susceptible to interference from background noise.

SPEECH RECOGNITION

Humans use not just auditory information in recognizing speech, but a host of non-verbal cues as well — more specifically, a speaker's facial movements (mouth, eyebrows, and so on), body