Early attempts to design systems for automatic speech recognition were mostly guided by the

theory of acoustic-phonetics, which describes the *phonetic elements* of speech (the basic sounds

of the language) and tries to explain how they are acoustically realized in a spoken utterance.

These elements include the phonemes and the corresponding place and manner of articulation

used to produce the sound in various phonetic contexts. For example, in order to produce a steady

vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates

through the vocal tract results in sound with natural modes of resonance similar to what occurs in

an acoustic tube. These natural modes of resonance, called the *formants* or *formant frequencies,*

are manifested as major regions of energy concentration in the speech power spectrum. In 1952,

Davis, Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition

for a single speaker [9], using the formant frequencies measured (or estimated) during vowel

regions of each digit. Figure 5 shows a block diagram of the digit recognizer developed by Davis

et al., and Figure 6 shows plots of the formant trajectories along the dimensions of the first and

the second formant frequencies for each of the ten digits, one-nine and oh, respectively. These

trajectories served as the “reference pattern” for determining the identity of an unknown digit

utterance as the best matching digit.

Early attempts to design systems for automatic speech recognition were mostly guided by the

theory of acoustic-phonetics, which describes the *phonetic elements* of speech (the basic sounds

of the language) and tries to explain how they are acoustically realized in a spoken utterance.

These elements include the phonemes and the corresponding place and manner of articulation

used to produce the sound in various phonetic contexts. For example, in order to produce a steady

vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates

through the vocal tract results in sound with natural modes of resonance similar to what occurs in

an acoustic tube. These natural modes of resonance, called the *formants* or *formant frequencies,*

are manifested as major regions of energy concentration in the speech power spectrum. In 1952,

Davis, Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition

for a single speaker [9], using the formant frequencies measured (or estimated) during vowel

regions of each digit. Figure 5 shows a block diagram of the digit recognizer developed by Davis

et al., and Figure 6 shows plots of the formant trajectories along the dimensions of the first and

the second formant frequencies for each of the ten digits, one-nine and oh, respectively. These

trajectories served as the “reference pattern” for determining the identity of an unknown digit

utterance as the best matching digit. The keyword spotting method (and its application in AT&T’s Voice Recognition Call Processing

(VRCP) System, as mentioned earlier), was introduced in response to the first factor while the second factor focused the attention of the research community on the area of dialog management.

Many applications and system demonstrations that recognized the importance of dialog

management over a system’s raw word recognition accuracy were introduced in the early 1990’s

with the goal of eventually creating a machine that really mimicked the communicating

capabilities of a human. Among these systems, Pegasus and Jupiter developed at the

Massachusetts Institute of Technology under Victor Zue were particularly noteworthy demos

[52,53], and the How May I Help You (HMIHY) system at AT&T developed by Al Gorin was an

equally noteworthy service that was introduced as part of AT&T Customer Care for their

Consumer Communications Services in 2000 [54].

Finite-state transducer