An Investigation of Audio-Visual Speech Recognition as Applied to Multimedia Speech Therapy Applications

Voula C. Georgopoulos
Department of Speech Therapy, Technological Educational Institute of Patras, Patras, GREECE
voulag@otenet.gr

Abstract

A multimedia speech therapy system should be able to be used for customized speech therapy for different problems and for different ages. The speech recognition must be designed to work with high inter- and intra-speaker variability. In addition to displaying text on a screen, recording the voice reading the text, analyzing the recorded spoken signal and performing speech recognition which includes identification of speech irregularities and tracking of patient progress, it should be capable of analyzing visual signal of the patients speech and provide visual as well as audio feedback. This implies that the synchronization of different media is important in realizing effective multimedia speech therapy applications. In order to perform speech recognition and identification tasks, time-frequency analysis and neural networks are proposed with integration of visual information.

1. Introduction

Multimedia-based tools for speech therapy rely heavily on speech recognition. A typical system displays text on a screen, allows recording of a voice reading the text, analyzes the recorded signal by comparing it to features of the "correct signal" in the data bank and performs a detailed speech recognition which includes speech error detection, such as prolonged syllables, gentle onsets, gentle transitions, and correct breathing. Visual cues are presented to the patient so that the patient can better understand what he/she must do to articulate the sound.

Integration of audio and visual integration is very important in speech perception in such systems. The patient must combine the information provided by the ear and eye for identification, recognition and then learning the production of the speech stimuli. A difficulty here is that the auditory and visual perception systems of the patient are not necessarily "normal". This means that there may be pathological problems of hearing and/or vision. The information may be integrated differently in patients with different problems. The goal is to provide adequate and appropriate visual and audio information to allow learning of the stimuli.

Modality can be defined as the perception via one of the perception-channels. In multimedia speech therapy applications we have trimodality since the user will look at the monitor, hear the sounds and move parts of the oral cavity etc. to reproduce the sound. In this paper we examine bimodality, i.e., auditory and visual. The multimedia system should be able to adequately produce human-like output (speech, faces, and movements).

Due to the variety of articulators (vocal cords, velum, tongue, lips, jaw, etc.), used by humans to produce speech, not all movements are visible. Similarly, due to periods of silence in a speech signal, it is not continuously audible. Thus, its components can be both visible and audible, only audible, or only visible [1].

To present visual information to the patient, the speech therapy system must first analyze the audio and visual signal of the patient doing the exercises and perform the speech recognition tasks. The speech recognition part must be designed to work with high inter- and intra- speaker variability. It is necessary to represent the signals in such a way that features can be extracted for recognition of phonemes and sounds. Representing the signal in a joint domain representation, (time-frequency domain) is needed because the frequency content of speech varies with time. The Wigner distribution and its moments, envelope, group delay and instantaneous frequency, are important quantities to show this frequency variation in time. They can be used as features to design a time-frequency based neural network system to provide patients with easy to interpret and use audio and visual information for speech improvement.

Section 2 discusses three communication disorders and appropriate multimedia speech therapy techniques along with the importance of visual information. Section 3 discusses speech sound characteristics. Section 4 presents analysis of speech in the time- frequency-domain and how this can be integrated with visual cues. Neural networks for speech recognition in speech therapy applications are presented in section 5 and section 6 gives a summary.

2. Representative communication disorders and the importance of visual information

Speech therapy techniques used for various speech and language disorders are widely varied. Here we present a few disorders where multimedia tools can be included in the speech therapy.

APHASIA: A language disorder caused by damage to the temporal lobe or higher up in the frontal lobe causing

0-7695-0253-9/99 $10.00 © 1999 IEEE

problems in understanding and/or formulating complex, meaningful elements of language. Areas that can be affected by aphasia include: 1) auditory comprehension 2) reading comprehension 3) oral expression 4) writing skills 5) ability finding the correct words. This is a severe communication impairment and so multimedia speech therapy techniques can focus on:

- imitating or repeating sounds and following commands. Sounds must be linked to pictures of objects so that language acquisition is improved.
- melotic therapy where there is an increase of language skills using music.
- Development of a multimedia communication board for patients to point to what they need and have the words be produced by the computer.

ARTICULATION DISORDERS: These are disorders in the production of individual speech sounds. Consonants are often misarticulated. There are four types of errors in articulation: substitution, omission, distortion, and addition.

Exercises that can be performed on a multimedia speech therapy system are: Sensory Perceptual Training where patients can hear their own mistakes using playback of recorded speech and comparison of their visual image of lip movements to animated articulation diagrams. With this method patients work on producing a new sound with cueing, on their ability to make sounds correctly in isolation, and their use of the sound in syllables, in words, in phrases and in sentences.

STUTTERING: Stuttering occurs when the forward flow of speech is interrupted abnormally. Typical dysfluencies/ stuttering are: repeating a sound, syllable, or part of a word, holding out a sound silently, etc. The exercises that can be used here are similar techniques as in articulation disorders, only the content will be different.

2.1. Visible Speech

Although speech perception is considered an auditory process, studies have shown that visual information provided by the movements of a talker's mouth and face strongly influences what an observer perceives even when the auditory signal is clear and unambiguous [2]. It is well known that lip-reading is necessary for the hearing impaired to partially understand speech. Even if the auditory modality is the most important for speech perception by normal hearers, the visual modality may allow subjects to better understand speech.

In a multimedia speech therapy environment, integration of visual and auditory information allows quicker learning since the various movements of the facial anatomy shown on the screen may be replicated by the patient. In [3] and [4] the relative intelligibility of stimuli uttered by a speaker in background noise for three modes of presentation: audio alone, audio plus lips and audio plus face were evaluated. It was shown the intelligibility scores for the "audio plus face" is the best of the three, whereas the "audio plus lips" is second best.

Synchronization and delays between auditory and visual information become a crucial issue. It is not possible to detect asynchrony between visual and auditory presentation of speech when the acoustic signal is presented less than 130 ms before or 260 ms after the continuous video display of the speakers face [5]. Since the correct production of speech by the patients relies on synchrony between the auditory and visual information, these numbers must be carefully observed.

Sometimes phonemes that are the easily audibly discriminable are difficult to distinguish visually, and vice versa. For instance, /p/, /b/, and /m/ look alike, although they sound different, and are often grouped together as one viseme. Speech recognizers often confuse /p/ and /k/, whereas they look very different on the speaker's lips. Thus, a synthetic face can improve the intelligibility of a speech synthesizer if facial movements are coherent with the acoustic flow that is supposed to be produced by them.

Since speech recognition is an important component of multimedia speech therapy systems, the next section deals with the general characteristics of speech.

3. Speech characteristics

Speech sounds can be classified as follows [6], [7]:

- vowels - are produced by exciting a fixed vocal tract shape with pulses of air from the vibration of the vocal cords. The vowel sound produced is determined by the position of the tongue, the jaw and the lips.
- diphthongs - are gliding monosyllabic speech sounds that start near the articulatory position for one vowel and move toward the position for another. They are produced by varying the vocal tract smoothly between appropriate vowel configurations.
- semivowels /w/ /y/ /l/ /r/ - are vowel-like sounds, characterized by a gliding transition in the vocal tract area between adjacent phonemes. The acoustic characteristics depend on the context.
- nasal constants /m/, /n/ - are sounds produced with glottal excitation and the vocal cord totally constricted at some point along the oral passageway.
- unvoiced fricatives /f/,/q/ /s/,/sh/- are produced by exciting the vocal tract by a steady air flow which becomes turbulent in the region of a constriction. The constriction determines which sound is produced.
- voiced fricatives /v/, /th/, /z/, /zh/- are the counterparts of unvoiced fricatives. The vocal cords are vibrating.
- voiced stops /b/, /d/, /g/ - are abrupt sounds produced by pressure build up behind a constriction and then released with pressure. The vocal cords vibrate.
- unvoiced stops /p/, /t/, and /k/- same as voiced stops without vibration of vocal cords.

Constrictions can be teeth, tongue, lips, glottis, etc. As an example, Table I [8] contains a summary of the timing and spectral characteristics for voiced/unvoiced stops and fricatives. It shows that temporal and spectral characteristics of each production mechanism vary. A

computer-based speech therapy system must use time and frequency features for speech recognition and speaker identification. Such features are timing of the preceding vowel, timing of constriction, timing of release, spectral content of constriction and spectral content of release. A time-frequency analysis tool that has been used for analysis of speech sounds is the Wigner distribution.

TABLE I. Features of Voiced/unvoiced contrast in stops and fricatives

Feature	*Voiced Stops*	*Voiced Fricatives*	*Unvoiced Stops*	*Unvoiced Fricatives*
Timing of Preceding vowel	Long preceding vowel	Long preceding vowel	Shortened preceding vowel	Shortened preceding vowel
Timing of Constriction	Brief oral closure	Brief oral constriction	Longer oral closure	Longer oral constriction
Timing of Release	Brief release transient 10-20 ms	No transient on release	Strong release transient & aspiration, 30-70 ms	No transient on release
Spectral constriction	Very low frequency sound during closure	very low fr. sound at constriction and correlated fluctuations in mid- and high- freq. regions	Silence	Strong mid- and high-frequency sound
Spectral Release	weak transient on release of closure, no aspiration	No transient on release	Strong transient on release	No transient on release

4. Analysis in time- and frequency-domains

The Wigner distribution (WD) is a quadratic-signal representation introduced in 1932 [9] and later used by as a tool for time-frequency analysis [10]. It is interpreted as signal energy density in time and frequency. The continuous WD of the analytic signal z(t) is defined by,

$$W(t,f)=\int_{-\infty}^{\infty} z\left(t+\frac{\tau}{2}\right)\cdot z^{*}\left(t-\frac{\tau}{2}\right)\cdot e^{-j2\pi f\tau}d\tau. \qquad (1)$$

For a discrete-time signal z[n], the discrete WD is,

$$W[n,f]=2\sum_{k=-\infty}^{\infty}\exp(-j2\pi kf)\cdot z[n+k]\cdot z^{*}[n-k]. \qquad (2)$$

To evaluate the DWD, a finite number of samples are used. Thus, the WD of a windowed signal is computed.

4.1. First-order moments of the WD

The definition of the instantaneous frequency is [11]

$$f_i(t)=\frac{1}{2\pi}\cdot\frac{d[\theta(t)]}{dt}, \qquad (3)$$

where $\theta(t)$ is the phase of the analytic signal, $z(t)=|z(t)|e^{j\theta(t)}$. The analytic signal of a real signal, x(t), consists of x(t) as its real part and the Hilbert transform as the imaginary part. Its Fourier spectrum has no negative frequencies. The instantaneous frequency shows the localization in time of the average frequency of a signal. It is also the first-order moment of the WD with respect to frequency:

$$f_i(t)=\frac{\int_{-\infty}^{\infty} fW(t,f)df}{\int_{-\infty}^{\infty} W(t,f)df}, \qquad (4)$$

where the integral in the denominator is equal to the envelope squared, or instantaneous power, given by

$$|z(t)|^2=\int_{-\infty}^{\infty} W(t,f)df. \qquad (5)$$

Group delay, τ_g is the delay of the envelope of the signal x(t) and is given by

$$\tau_g(f)=-\frac{1}{2\pi}\cdot\frac{d[\phi(f)]}{df}, \qquad (6)$$

where $\phi(f)$ is the phase of the Fourier transform of x(t). The group delay is also the first-order moment of the Wigner Distribution with respect to time:

$$\tau_g(f)=\frac{\int_{-\infty}^{\infty} tW(t,f)dt}{\int_{-\infty}^{\infty} W(t,f)dt}, \qquad (7)$$

where the integral in the denominator is the energy spectral density, given by

$$|Z(f)|^2=\int_{-\infty}^{\infty} W(t,f)dt. \qquad (8)$$

These quantities reveal the frequency content of the speech signal, when a particular frequency appears in time, what the instantaneous power is of the speech signal at a given time instant, and what the mean frequency is at a given time instant. In the next subsection shows the usefulness of the these quantities.

4.2 Important time-frequency characteristics

Time-frequency quantities important to speech are the envelope, group delay and instantaneous frequency:

- envelope - It shows location of bursts of energy and transitions from high energy to low energy. It can be used to segment individual phonemes in speech.
- group delay - Intuitively, it shows at which instant a given frequency appears. Resonant frequencies of vocal tracts are peaks of smoothed group delay.

- instantaneous frequency - is the frequency of a signal at a given instant in the average sense. When smoothed, it shows variations of average frequency within a speech token. Also, the pitch of the signal can be obtained indentifying individual speakers.

These quantities in conjunction with the visual signal, can provide phoneme segmentation and audio/visual synchronization. An example of a speech signal as a function of time is shown in Fig. 1.

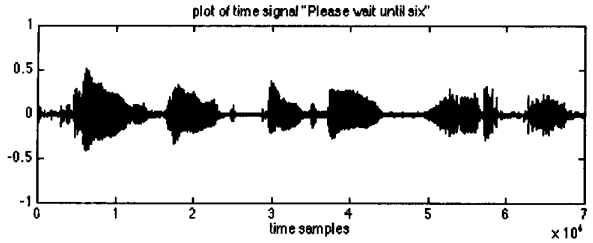

Fig. 1: Example of a speech signal as a function of time by a female speaker saying "Please wait until six".

The time-frequency plot (smoothed WD) of the speech signal (Fig. 2) reveals the high frequency noise-like characteristics of the sound of s in 'please' and the sounds of s and x in 'six'. The various vowels have different spectral characteristics depending on their formants. The frequency spread of the vowels is much less than that of the fricative sounds.

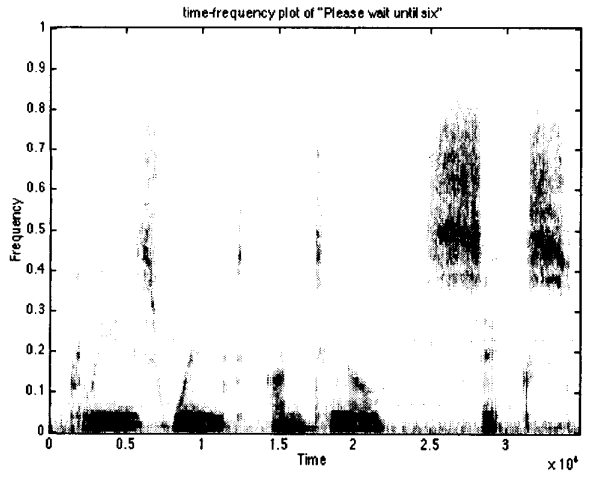

Fig. 2: Time-frequency plot (smoothed Wigner distribution of the speech signal of Fig. 1.

From the smoothed WD we obtain the estimate of the envelope showing high and low energy variations as a function of time (Fig. 3a). The envelope can be used to isolate individual words and phonemes. An estimate of instantaneous frequency shows the mean frequency at a given time instant (Fig. 3b).

In order to obtain the group delay it is important to separate the individual phonemes. Figure 4a shows a plot of the sound /x/ from the word "six". The peaks of the lowpass filtered group delay shown in Fig. 4b indicate the location in time location of the dominant frequencies in the phoneme. For vowels, these positions indicate the formant locations for vowels. Formants are the frequencies of resonances of the vocal tract which uniquely identify a vowel.

Once the signal features of interest are determined, a series of neural networks can be used to perform speech recognition and to provide the patients with audible and visible feedback on their progress.

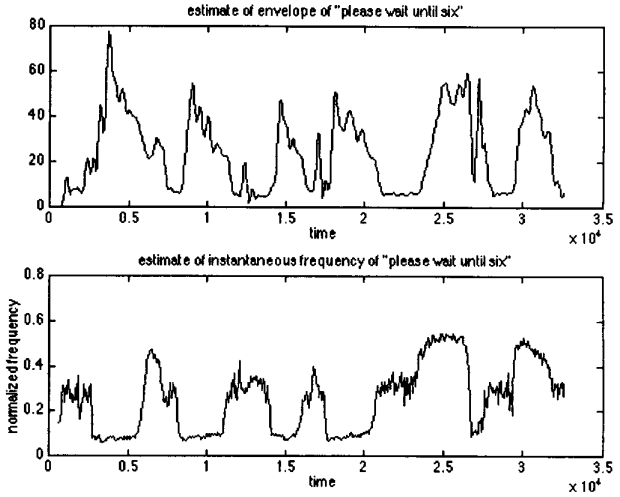

Fig. 3a and 3b: Envelope and Instantaneous frequency plots of speech signal in Fig.1.

5. Neural networks for audio-visual speech recognition in speech therapy applications

Speaker identification is needed for individual tracking of patient progress as well as speech recognition so that the system recognizes the spoken words and separates between regional accents, speaker age and gender or speech problems. Neural networks have become very popular in both areas since they require weaker assumptions about the statistical properties of the input data than more traditional signal processing techniques for speech recognition and speaker identification. Important neural net properties here are:

- ability to learn: A speech therapy computer system needs to be able to learn both the actual word and its special features in the time-frequency plane. It also, needs to learn specific patterns of a patient's speech so that monitoring of progress is achieved.
- robustness: Neural networks by their design allow for noisy inputs. This is important since speech signals are inherently noisy. Also, in a speech therapy system it is necessary to allow for speaker variability (age, gender, regional accents). However, the variability cannot be too high to isolate speech problems.
- parallelism: Because neural networks are inherently parallel in nature, they can process the speech signals fast. This is necessary because feedback must be fast, so that patient does not have to wait.

- generalization: Networks learn the underlying patterns of speech so they can generalize from speech used for training to new examples of speech. This is essential since two speech signals are never the same [12].

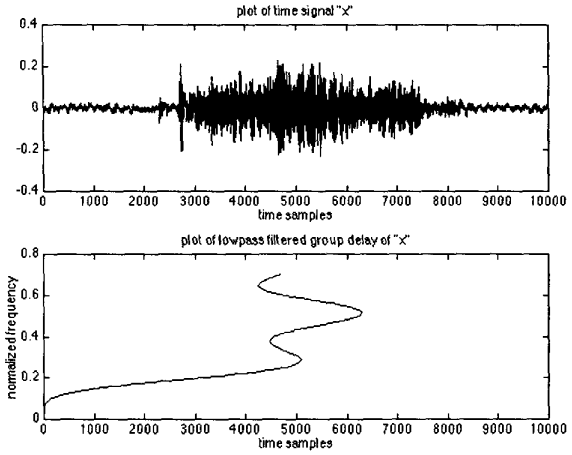

Fig. 4a and 4b: Segment of speech signal corresponding to /x/ sound and it's group delay.

Huang and Lippmann [13] demonstrated that neural networks can form elaborate decision surfaces from speech data. They applied a multilayer perceptron to a collection of vowels produced by men, women and children, using the first two formants of the vowels as the input representation. After 50,000 iterations of training using backpropagation, the network produced the decision regions separating the sounds HOD, WHO'D, HAWED, HEED, HID, HEAD, HAD, HOOD, HUD, HEARD.

For a computer based speech therapy system, we want to train a network with many speakers of different ages (native speakers of the language of interest without known speech problems). Thus, a feature space will be defined in the time-frequency domain. A similar network to the Lippmann network then should be developed, but instead of only using the first two formants as input, the envelope, instantaneous frequency and group delay should be included. Then a second network must be trained for deviations of envelope, instantaneous frequency and group delay for known speech problems.

Appropriate visual parameters must be used in a visual display of the production of the speech sound for better comprehension of the production method and thus, a successful multimedia speech therapy tool. For a speech therapy terminal, in addition to the actual face, an animated articulation diagram should be included showing the position of tongue, glottis, etc. Visual displays allow patients to compare their own voice and articulation patterns with the pre-defined stored ones.

To compare the articulation of the lips to the "correct" one, the visual signal of the patient's lips must be analyzed. In Figure 5 the points x1 - x10 of the outline of the patient's lip are used in addition to the parameters openness of mouth, width of mouth and movement of lower lip to determine the actual sound spoken. It is difficult to view the position of the tongue and teeth in most cases so these are not analyzed. A reference measurement is taken first in order to identify the subject's individual features of the lips.

Fig. 5: Lip shape extracted from video signal along with its outline and the 10 points used as reference.

Figure 6 shows a proposed block diagram of a multimedia speech recognition and identification system for speech therapy applications. The Speech Utterance goes through a signal processing block where time-frequency analysis is performed. Then estimates of the envelope, instantaneous frequency and group delay are obtained. These are used to extract features of importance which segment the individual phonemes and help classify the sound into what category of speech sound it belongs. The recognition is the last part of the first process where a comparison is made to phonemes making up words in a database. The second part of the computer based speech therapy system consists of a comparison to the desired sound and a comparison to the patient's progress using the patient database.

The video of the patient's face is image processed so that the lip area is isolated. It is then synchronized with the audio information using the envelope and instantaneous frequency information so that individual phonemes are resolved. For each individual phoneme, based on the measurements taken of the lips points mentioned earlier and information on the instantaneous frequency and group delay, a viseme is produced. This viseme is compared with the appropriate viseme stored in the computer for the phoneme and appropriate animations show the patient the difference between what he/she said and what he/she should've said.

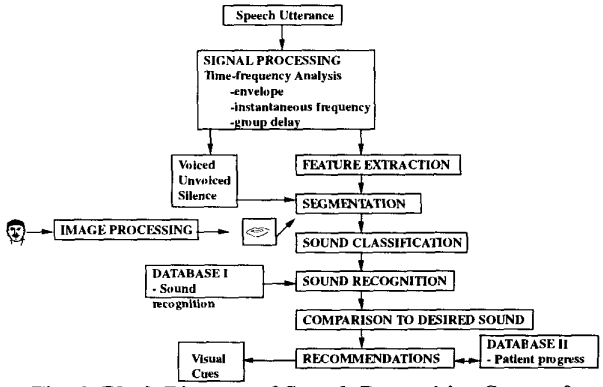

Fig. 6: Block Diagram of Speech Recognition System for Speech Therapy.

What is currently being investigated is how to train a neural network with time-frequency inputs from group delay, envelope and instantaneous frequency and perhaps other inputs for a phoneme sound and obtain a range of values for outputs corresponding to configurations of the

oral cavity.Output characteristics may include lip closure, tongue placement, placement of teeth, burst duration and mouth articulation for vowels. Each range of values can correspond to a different "configuration" in the mouth cavity. Figure 7 shows such a neural network diagram.

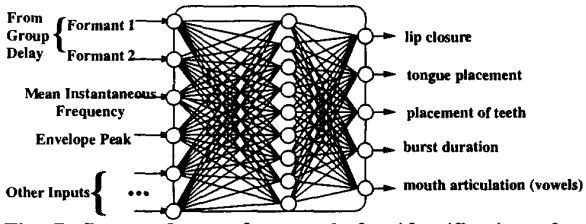

Fig. 7: Proposed neural network for identification of speech problems including time-frequency information.

The output of the neural network must be presented to a patient in a useful format so that it is easy to understand. One way of presenting this information to the patient could be in terms of the mouth articulation diagrams [14]. For example, for each range of values for that particular output of the neural network could correspond to a different articulation diagram.

6. Summary

The integration of the various modalities, auditory, visual and motion are very important in the speech recognition, perception and reproduction which are necessary in an effective multimedia speech therapy application. In this paper we discussed the issues surrounding audio-visual integration for speech recognition in multimedia speech therapy applications.

The need for analysis in the time-frequency domain was explained. The time-frequency components of interest are: the envelope, group delay, and instantaneous frequency. From these quantities along with the visual information, one can perform phoneme segmentation, estimate first two formants, and obtain pitch information. An example of a speech signal analysis in time-frequency domain was presented where the smoothed Wigner distribution, envelope, group delay, and instantaneous frequency were shown. The properties of neural networks that are important for this type of speech recognition problem were discussed.

Finally, a proposed diagram of a multimedia speech recognition system based on neural networks and time-frequency analysis for speech therapy applications was presented including individual patient progress identification.

7. References

[1] L. Schomaker, et al., "A Taxonomy of Multimodal Interaction in the Human Information Processing System," Report of the Esprit Project 8579 MIAMI, February, 1995.

[2] Kerry P. Green, "The Use of Auditory and Visual Information in Phonetic Perception," Speechreading by Man and Machine: Models, Systems and Applications, NATO Advanced Study Institute 940584, August 28 to September 8, 1995, Château de Bonas, France.

[3]. C. Benoît, T. Mohamadi, and S. Kandel, "Effects of phonetic context on audio-visual intelligibility of French speech in noise," J. of Speech & Hearing Research, 1994.

[4] B. Le Goff, T. Guiard-Marigny, and C. Benoît, "Read my lips ... and my jaw! How intelligible are the components of a speaker's face?", Proceedings of the 4th Eurospeech Conference, Vol. 1, 291-294, Madrid, Spain.

[5]. B. Dodd and R. Campbell, editors, Hearing by Eye: The Psychology of Lip-reading, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[6]. L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[7]. W. H. Perkins and R. D. Kent, Functional Anatomy of Speech, Language and Hearing: A Primer, Allyn and Bacon, Boston, 1986.

[8]. J. M. Pickett, The Sounds of Speech Communication, University Park Press, Baltimore, MD, 1980.

[9]. E. P. Wigner, "On the Quantum Correction for Thermodynamic Equilibrium," Phys. Rev., 1932. vol. 40: pp. 749-759.

[10]. T.A.C.M. Claasen and W.F.G. Mecklenbräuker, "The Wigner Distribution - A Tool for Time- Frequency Analysis Parts I-III," *Phillips J. Research*, 1980. vol. 35: pp. 217-250, 276-300, 372-389.

[11]. J. Ville, "Theorie et Application de la Notion de Signal Analytique," *Cables et Transmissions*, 1948. vol. 2A(1): pp. 61-74.

[12]. J. Tebelskis, Speech Recognition Using Neural Networks, Ph.D. dissertation, Carnegie Mellon University, May 1995.

[13]. W. M. Huang and R. Lippmann, "Neural Net and Traditional Classifiers," in Neural Information Processing Systems, D. Andersen (ed.), 387-396. New York: American Institute of Physics.

[14]. J. L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd ed., Springer-Verlag, New York, 1972.