A computer based speech therapy system for articulation disorders

Voula C. Georgopoulos[1], Georgia A. Malandraki[1] and Chrysostomos D. Stylios[2]

[1]Dept. of Speech Therapy, Technological Educational Institute of Patras, Greece
[2]Dept. of Electrical and Computer Engineering, University of Patras, Greece

ABSTRACT: A computer speech therapy system for articulation disorders 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 providing 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. The computer based speech therapy system under development is for Greek language.

INTRODUCTION

Articulation is the process by which sounds, syllables, and words are formed when articulators, i.e. tongue, jaw, teeth, lips, and palate alter the air stream coming from the vocal cords. Due to the fact that the correct production of speech is dependent on different factors, articulation problems may frequently occur. An articulation problem appears when a person produces sounds, syllables or words incorrectly so that listeners do not understand what is being said or they have to pay more attention to the way the words sound than to what they mean. Most articulation errors fall into one of three categories: omissions, substitutions, or distortions [1].

The treatment of articulation disorders lies in the expertise of speech and language therapy. Treatment is critical if one considers the possible impact of an articulation problem on one's social, emotional, educational, and/or vocational status. It is widely known that speech is the most important means of communication and thus, the quality of our lives is affected by the adequacy of it. Therefore, the goal of this work is to design a computer based speech therapy system that will contribute to the treatment of articulation disorders of Greek language.

The under-development computer based speech therapy system provides target speech sounds, allows recording of a patient speaking, analyzes the recorded signal by comparing it to the features of sounds in the data bank and performs a detailed speech recognition. Visual cues (animations) are presented to the patient in order to see how he/she produced each sound and better understand what he/she must do to correctly articulate the sound. All speech sounds that can be produced are included in the International Phonetic Alphabet (IPA), a standard set of symbols for transcribing the sounds of spoken languages [2]. The characteristics of these sounds will be stored in the data bank of this system.

Due to the variety of articulators (vocal cords, velum, tongue, lips, jaw, etc.), that have to cooperate in order for humans to produce speech, not all movements of speech production are visible. Similarly, there are periods of silence in a speech signal and so it is not continuously audible. Thus, the components of speech can be visible and audible, only audible, or only visible. Therefore, animations along with sound feedback are instrumental in the speech therapy and make apparent the need of an integrated multimedia system.

To present visual information to the patient, the speech therapy system must first analyze the audio 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.

Neural Networks have been considered as a great problem solution for speech processing and recognition. NN have many interesting characteristics and abilities: parallel computation, possessing robustness and fault tolerance, adaptive learning ability and approximation of any nonlinear dynamical system.

The next section discusses what articulation is and it is followed by a discussion of the International Phonetic Alphabet and speech sound characteristics. A presentation of the analysis of speech in the time- frequency-domain and how this can be integrated with visual cues is provided. Necessary neural network techniques for speech recognition in speech therapy applications are then discussed and finally, a summary is provided.

DEFINITION OF ARTICULATION DISORDERS

The ability that we have to produce speech sounds is described as articulation. Specifically, articulation is a general term used in phonetics to denote the physiological movements involved in modifying the airflow, in the vocal tract above the larynx, to produce the various speech sounds [3].

Any problem related especially to speech sound production is considered an articulation disorder. The mechanism of articulation has to do with the coordination of many muscles,

including those of breathing, the vocal cords, tongue, and soft palate. Even the size and the shape of teeth, tongue, and hard and soft palates have an important part in the articulation system. If any one of these elements is impaired, an articulation disorder may result [4].

Specifically, the term articulation disorder refers to omissions, distortions or substitutions of sounds in speech. In a typical substitution error, for example, a child may say /γ/ against of /ʃ/ in the Greek word /ʃaso/ so it would be heard as / γaso/. These kinds of mistakes are systematic which means that a child may only misarticulate a couple of sounds, but he/she does so in all words that contain those sounds. In many cases that results in an unintelligent speech while in others the speech remains intelligible which is a fact that depends on the frequency of the misarticulated sounds. In any of these cases the articulation disorder constitutes a problem for the patient that must be solved. From the clinical experience found, a few of the most common articulation errors that Greek children make are shown in Table I.

Table I. Some of the Common Articulation Errors in Greek

Type of error	Target sound	Produced sound
Substitution	ʃ	/γ/
Substitution	/s/	/ç/ or ʃ
Substitution	/v/	/f/
Distortion	/ð/	/θ/

The main factors that can contribute to the existence of an articulation disorder are described as: Organic problems (e.g. cleft palate, dental abnormalities, mental retardation, etc), Apraxia of speech (mainly associated with brain injury), Functional problems (e.g. a bilingual child), Tongue thrust (immature swallow and pressure of the tongue against the front teeth that causes misalignment of the teeth), and Developmental delay, which is the cause of most articulation disorders.

A speech therapist in order to be completely aware of any articulation disorder must have a deep knowledge of what exactly is going on when an articulation disorder takes place.

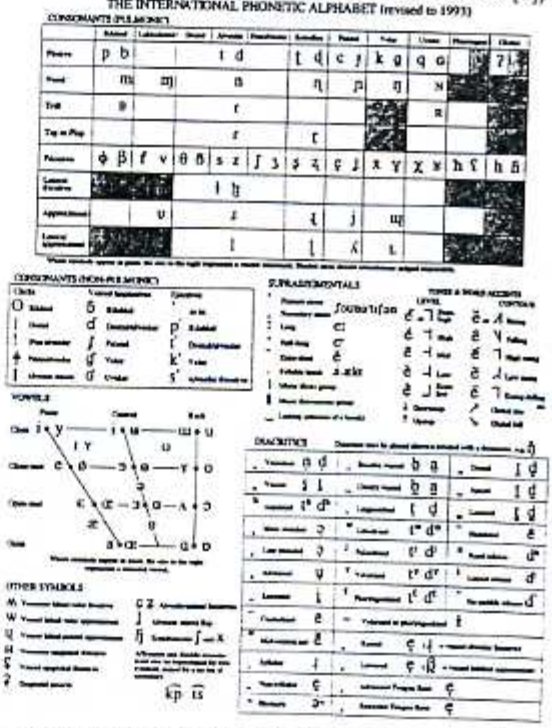
THE INTERNATIONAL PHONETIC ALPHABET

The International Phonetic Alphabet (IPA) is a means that provides that kind of knowledge to the speech therapist because it constitutes an alphabet, which contains symbols of all the sounds that can be produced by human beings. It primarily uses Roman characters; other letters are borrowed from different scripts (e.g., Greek) and are modified to conform to Roman style. Diacritics are used for fine distinctions in sounds and to show nasalization of vowels, length, stress, and tones [5].

In the IPA, sounds are classified according to their place and manner of articulation in the vocal mechanism [3]. By the term place of articulation ,we mean the place where the articulators come into contact or where the flow of the air is more constricted. Additionally, by the term manner of articulation, we mean basically the way that the air exits the mouth or nose [6].

According to the IPA, the speech sounds are divided into subcategories of the manner and of the place of articulation. Table II shows analytically these categories and the sounds that belong to each one of them.

TABLE II (from the International Phonetic Association [7])



The subcategories for the place of articulation are (adapted from [8], [9], and [10]):

- Bilabial sounds: the sounds that are being produced by the contact of the lips.
- Labiodental sounds: the sounds that are being produced by the contact of the upper teeth with the lower lip.
- Dental sounds: these sounds are produced by the contact of the tip of the tongue with the upper teeth.
- Alveolar sounds: they are the sounds that are produced by the contact of the tip of the tongue with the alveolar ridge.
- Postalveolar sounds: they are the sounds that are produced by the contact of the tip or the blade of the tongue with the spot that lies immediately behind the alveolar ridge.
- Retroflex sounds: they are the sounds that are produced by the contact of the lower part of the tongue with the spot that is slightly behind the previously mentioned spot.
- Palatal sounds: they are the sounds that are produced by the contact of the blade of the tongue with the hard palate.
- Velar sounds: they are the sounds that are produced by the contact of the back part of the tongue with the soft palate.
- Uvular sounds: they are the sounds that are produced by the contact of the back part of the tongue with the velum.
- Pharyngeal sounds: they are the sounds that are produced by the base of the tongue and the pharyngeal walls.
- Glottal sounds: they are the sounds that are produced by the constriction or the closure of the vocal cords.

The subcategories for the manner of articulation are (adapted from [8], [9], and [10]):

- Plosives: they are produced by a sudden release of the air which is slightly heard like a small explosion.

- Nasals: they are produced while the air exits through the nose.
- Trills: they are produced by quick rhythmic movements or vibrations of the articulator.
- Flaps or taps: they are produced by a very brief contact of the articulators.
- Fricatives: they are produced by a narrow closure of the articulators that enables the air to exit with some friction.
- Lateral fricatives: they are produced with the same way as the previous sounds but here the air exits through the lateral sides of the articulators.
- Semivowels or approximants: they are produced by an approach of the articulators that is more narrow than that for vowels but wider than that for the production of fricatives.
- Lateral approximants: they are produced the same way as the previous sounds but here the air exits through the lateral sides of the articulators.
- Affricatives: they consist of two symbols because we have two sounds being heard, one plosive and one fricative or semivowel.

Also the sounds are divided to voiced and voiceless sounds according to whether the vocal cords are vibrating or not during their production. When the vocal cords are vibrating during the production of a sound, then this sound is called a voiced one, while when the vocal cords are not vibrating during the production of a sound, then that sound is called a voiceless sound.

INTEGRATED COMPUTER BASED SYSTEM FOR SPEECH THERAPY

Despite the benefits of the IPA for the speech therapist, the patient, who is more likely not to be aware of this alphabet, cannot have a feedback for his/her mistakes. Also when used in printing and typing it utilizes of a large number of special symbols in addition to the letters of the Roman alphabet that constitute its core. So our proposed computer program will provide a "translation" of this alphabet to a realistic picture of the patient articulators and that will be the best visual feedback of himself/herself that he/she could have in order to understand and later correct the existing articulation errors.

To present visual information to the patient, the speech therapy system must analyze the audio signal of the patient speaking and perform the speech recognition tasks. 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 time-frequency domain is needed because the frequency content of speech varies with time. The next section discusses a joint time-frequency representation.

ANALYSIS IN TIME- AND FREQUENCY-DOMAINS

The Wigner distribution (WD) is a quadratic-signal representation introduced in 1932 [11] and later used by as a tool for time-frequency analysis [12]. It is interpreted as signal energy density in time and frequency. The continuous WD of the analytic signal z(t) is defined by,

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

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

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

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

FIRST-ORDER MOMENTS OF THE WD

The definition of the instantaneous frequency is [13]

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

where θ(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:

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

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

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

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

$$\tau_g(f)=\frac{1}{2\pi}\frac{d[\varphi(f)]}{df} \quad (6)$$

where φ(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:

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

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

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

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 [14]. 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:

- envelope - the square root of instantaneous power. It shows location of bursts of energy and transitions from high energy to low energy. It can be used to segment individual phonemes in a speech signal.

- group delay - Group delay at a specific frequency was defined and measured as the time taken for the envelope of a narrow-band signal centered at that frequency to propagate through the system under test. Intuitively, it shows at which instant in time does a given frequency appear. One can extract resonant frequencies (formants) of vocal tracts from peaks of smoothed group delay.
- instantaneous frequency - the frequency of a signal at a given instant in time in the average sense. When smoothed to remove rapid variations, it shows the variations of average or carrier frequency within the speech token. After low-pass filtering the pitch of the speech signal can be obtained.

Combinations of all three quantities can be used for obtaining timing and frequency characterization of a speech signal.

EXAMPLES

As examples showing the differences in the time frequency domain of misarticulated sounds, four signals were chosen: two with correct target consonant sounds and two with the respective misarticulated consonant sounds. The target pseudowords used were /ava/ and /asa/ and the respective misarticulated ones are /afa/ and /aça/.

First we will discuss the differences in the articulation of each of the sounds [15]:

/f/ Voiceless labiodental fricative. The lower lip is brought close to the upper teeth, occasionally even grazing the teeth with its outer surface, or with its inner surface (figure 1).

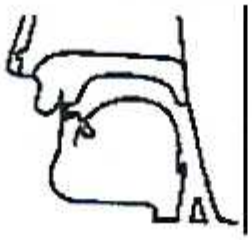


Figure 1. Articulation diagram for /f/ and /v/

/v/ Voiced labiodental fricative. Same as above, but with vibration of the vocal cords.

/s/ Voiceless alveolar fricative. Produced by bringing the end of the tongue close to the alveolar ridge (figure 2).

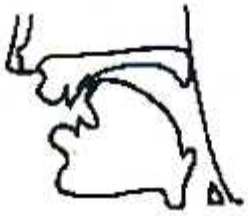


Figure 2. Articulation diagram for /s/

/ç/ Voiceless palatal fricative. The tongue body forms a groove and approaches the hard palate (figure 3).

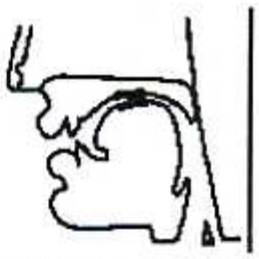


Figure 3 Articulation diagram for /ç/

In the articulation of the /v/ sound a frequency error is the absence of vibration of the vocal cords which leads to the erroneous /f/ sound.

In the present example of the misarticulated /s/ sound leading to the /ç/ sounds, this is caused by the position and shape of the tongue.

The Wigner distribution of /ava/, /afa/, /asa/, and /aça/ are shown in Figures 4,5,6, an 7 respectively (center plot of each figure) along with the signal in the time domain (bottom plot of each figure) and the spectrum of the signal (left plot of each figure). In the top, right corner and right plots of each figure, respectively, are shown the envelope, instantaneous frequency, and group delay for the sounds that correspond to the consonants only.

It is clear that the signal in the time domain and in the frequency domain alone does not provide sufficient information on the differences between the target sounds and misarticulated ones. Differences between target sounds and produced consonant sounds are apparent in the Wigner distribution, but since the consonant sounds are of much lower energy levels than the vowel sounds, much of the detail is not present. The zero and first order moments, envelope, group delay and instantaneous frequency of the consonant sounds, make the differences apparent.

Thus, they can be used as feature vectors for a neural network scheme that will categorize the sounds in terms of place and manner of articulation as well as whether they are voiced or unvoiced.

PROCESSING OF AUDIO SIGNAL USING NEURAL NETWORKS

Neural networks are very popular in speech recognition and speaker identification since they require weaker assumptions about the statistical properties of the input data than more traditional signal processing techniques. Properties of importance 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.
- parallelism: Because neural networks are inherently parallel in nature, they can process the speech signals fast. This is necessary because feedback must be timely.

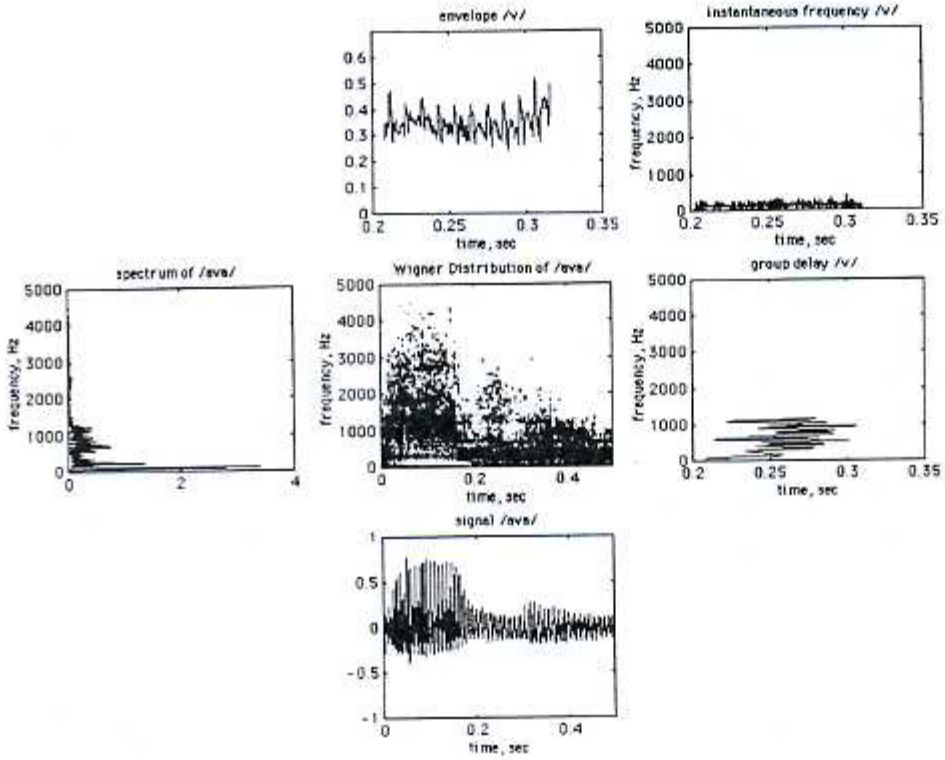


Figure 4.Analysis in the time-frequency plane of the pseudoword /ava/ and the target consonant sound /v/.

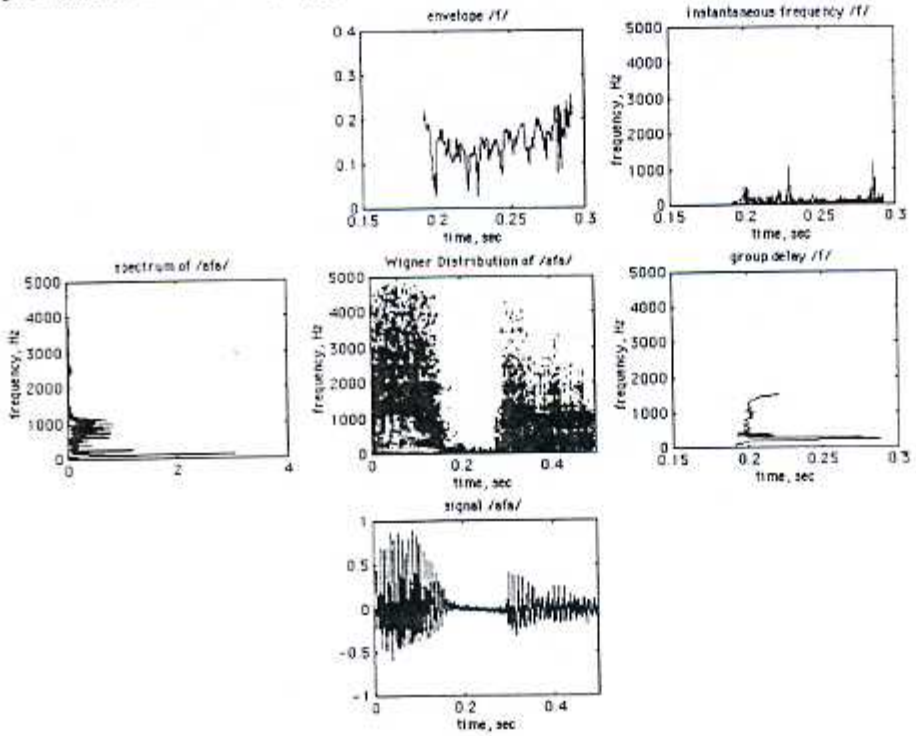


Figure 5.Analysis in the time-frequency plane of the pseudoword /afa/ and the produced consonant sound /f/ (misarticulated /v/).

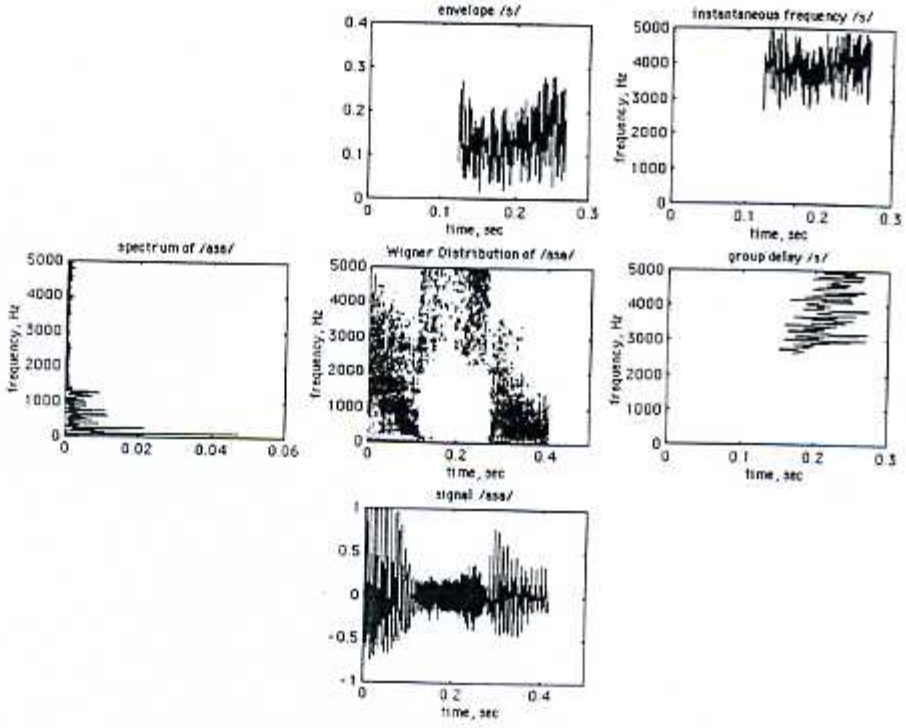


Figure 6.Analysis in the time-frequency plane of the pseudoword /asa/ and the target consonant sound /s/.

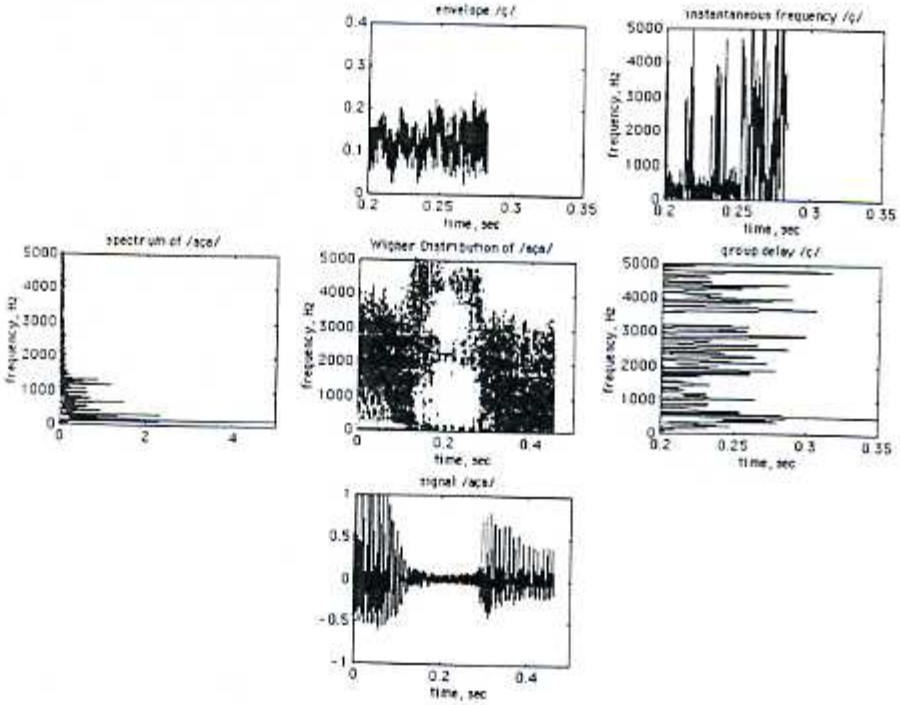


Figure 7.Analysis in the time-frequency plane of the pseudoword /aça/ and the produced consonant sound /ç/ (misarticulated /s/).

- 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 [16].

Huang and Lippmann first demonstrated that neural networks can form elaborate decision surfaces from speech data (for vowels) using a multilayer perceptron and the first two formants of the vowels as the input representation [17].

What is currently being investigated is how to train a neural network system 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 jaw, burst duration, and vibration of vocal cords. Clearly, a single neural network cannot perform such a complex function. Separate neural networks will be used for each output characteristic leading to a particular value. The combinations of values for the characteristics can lead to a different articulation configuration. Figure 8 shows such a *conceptual* neural network diagram, where the neural network in the figure is a combination of neural networks for each output.

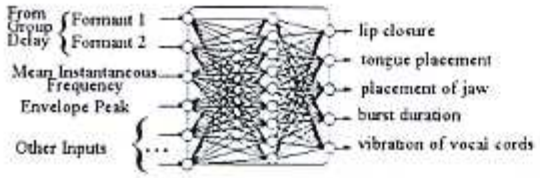


Figure 8. Proposed *conceptual* neural network for identification of speech problems including time-frequency information. This corresponds to a group of neural networks for each output.

For signals, such as those shown in the examples above, the segmentation of the consonants from the vowels is probably the first necessary step of processing of the time-frequency signal. This can be done using a low pass-filtered version of the envelope in combination with an automatic syllabic detection algorithm such as that of Melmerstein [18]. With this, peaks and dips are found in the envelope waveform while at the same time duration and absolute level constraints are placed as well [19].

Since the size of the envelope, instantaneous frequency, and group delay vectors may be too large to effectively be used by neural network classification algorithms, feature extraction is necessary to produce feature vectors from these.

A technique that can be used for feature extraction is the Intra- and Inter-Class Separability of the Feature Space [20]. To evaluate the scene discrimination capability of the selected features, the intra-class and inter-class scattering matrices are calculated [21]. The intra-class scattering matrix reveals the scattering of samples around their respective class centroids, and is defined by

$$S_{intra} = \sum_{i=1}^{N} P(\omega_i) E\left\{ (X - M_i)(X - M_i)^T \mid \omega_i \right\}. \qquad (9)$$

where, $P(\omega_i)$ is the a priori probability of class ω_i, X is the sample feature vector, M_i is the mean feature vector (centroid) of class ω_i, N is the number of classes. On the other hand, the inter-class scattering matrix is defined as:

$$S_{inter} = \sum_{i=1}^{N} P(\omega_i)(M_i - M_0)(M_i - M_0)^T. \qquad (10)$$

where $M_0 = \sum_{i=1}^{N} P(\omega_i) M_i$. (11)

The diagonal items in these two matrices characterize the intra- and inter-class separability of individual features. If the diagonal item in the intra-class scattering matrix is small while that in the inter-class matrix is large, then the corresponding feature has good class separability. The off-diagonal items in these two matrices reveal the correlation between different features. These measures can be used to eliminate highly correlated features and reduce the dimensionality of the feature space [20].

Neural Network Classifier Feedforward neural networks have been used successfully as pattern classifiers in many applications. Conventional multilayer perceptron (MLP) use the all-class-in-one-network (ACON) structure. But such network structures have the burden of having to simultaneously satisfy all the desired outputs for all classes, so the number of hidden units tends to be large [20]. Two other classification techniques are under investigation. The first is the one-class-in-one-network (OCON) structure, where one subnet is designated for recognizing one class only [22]. In this structure, each subnet is trained individually using the back-propagation algorithm so that its output is close to 1 if the input pattern belongs to this class, otherwise the output is close to 0. Given an input feature vector, such as those produced from the envelope, group delay or instantaneous frequency, it is classified to the class whose subnet gives the highest score. An advantage of the OCON structure is that one can accommodate a new class easily by adding a subnet trained for this class.

Another classification neural network is the Learning Vector Quantization Classifier [23] which has one competitive layer followed by a linear layer. The first layer classifies incoming vectors into subclasses, whereas the second one merges the final ones into final target classes [24]. This type of network is useful when the classes of the patterns are not linearly separable in the pattern space. The network is trained by the LVQ2 algorithm. During classification, the input vectors are classified according to the 1-Nearest Neighbor rule.

The output of the neural network must then 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 [25]. For example, for each range of values for that particular output of the neural network could correspond to a different animated articulation diagram.

Figure 9 shows a proposed block diagram of a multimedia speech recognition and identification system for speech therapy of articulation disorders. The Speech Utterance passes 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