9 1

AUTOMATIC SPEECH QUALITY ANALYSIS

WITH APPLICATION TO SPEECH TRAINING

by

ROLF EXNER

B.E.(Hons.), University of Tasmania, 1977
B.Sc., University of Tasmania, 1975

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Department of Electrical Engineering)

We accept this thesis as conforming
to the required standard.

THE UNIVERSITY OF BRITISH COLUMBIA

August, 1979

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Rolf Exner

Department of Electrical Engineering

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

1979 August 23

ABSTRACT

A number of aspects of speech training involve assessing the quality of the student's speech. It is of interest to determine whether such speech quality analysis can be done automatically. This thesis provides a preliminary answer to that question by proposing and then evaluating a set of quality measures for comparing the quality of two segments of speech.

Speech quality is taken to be the lack of defects in the articulatory and prosodic components of speech. It is a non-quantitative definition from speech pathology that can meet the needs of speech training. Speech defects common among deaf children and students of English as a second language are reviewed, and classified according to this scheme.

The speech quality measures are based on a linear prediction model of speech, and adapt several techniques from the field of speech recognition. Evaluations using speech with known quality defects show that the articulatory measures are effective in detecting most of the common errors of articulation, with the exception of ones between nasal sounds. The prosodic quality measures of loudness and timing give very useful indications of syllable stress and voicing errors. The timing measure is derived from the optimal time-warping curve between the two utterances, and provides an accurate means of tracking speed variations in speech. Differences between speakers tend to mask articulatory quality errors, but have little effect on the prosodic quality measures. An articulatory distance measure is proposed that partly counters these interspeaker differences.

Work remains to be done in a number of key areas, but the results of this preliminary investigation suggest that automatic speech quality analysis by computer is practical and may one day become a versatile tool for speech training.

TABLE OF CONTENTS

page

LIST OF FIGURES

ACKNOWLEDGEMENT

It is my pleasure to acknowledge the generous assistance I obtained from many quarters over the past two years. I would especially like to thank my supervisor, Dr M.R. Ito, for his interest and guidance throughout the work, and my wife Heidi, for her unfailing help and encouragement. I am grateful to Mr D. Laplante and the U.B.C. Department of Oceanography for making available the digitizing facility, and to my fellow students for contributing the by-now-much-analyzed speech data.

Financial assistance was provided through a Canadian Commonwealth Scholarship 1977-1979, and a U.B.C. Teaching Assistantship, 1978-1979. Both are gratefully acknowledged.

CHAPTER 1

INTRODUCTION

1.1 WHY SPEECH QUALITY ANALYSIS

There has been considerable interest in recent years in the development of technical aids for the purpose of speech training. These aids, which in general depend on a visual display of certain features of speech, are intended for use in the classroom by deaf children and language students, and possibly by others. Children with a severe hearing impairment of early onset face an enormously crippling social handicap if they are unable to learn useful speech, and the difficulties of teaching them by traditional means have provided a strong incentive for research into improved methods of training.

Unfortunately, speech training aids have had only limited success, and the need to investigate and design better aids remains as strong as ever. One difficulty has been that the devices were all intended to help with specialized aspects only of the speech learning problem. This thesis describes a new tool for use in speech training that is general in approach and applicable to many facets of the problem. It makes a direct evaluation of the quality of the student's speech from a comparison between it and the teacher's speech. Indications are that a wide range of quality errors can be detected by the method, and diagnostic information, i.e. information as to how and why the speech is defective, is additionally available. The speech quality analysis system is implemented on a computer via linear prediction methods, and borrows from techniques developed for speech recognition and speaker identification. It will likely find most use in those speech training applications in which an extra feedback channel can benefit the

learning process. This is the case with deaf children who lack the normal auditory capacity for comparing their attempts at speech with those of others around them; it is also the case with students learning another language, who through long exposure to their native tongue have lost some of their ability to make fine auditory discriminations with the foreign sounds being learned.

The next section examines more closely aids for speech training, including their advantages, limitations, and current capabilities. It will be apparent from the discussion that the design and implementation of a speech training aid is a major interdisciplinary undertaking. Therefore, this thesis can be concerned only with an investigation of the feasibility of the quality analysis approach, and not with the construction of a ready-to-use speech training aid.

1.2 AIDS FOR SPEECH TRAINING

Speech training aids are the result of applying technology to the difficult problem of teaching speech, and as such they potentially have many advantages over traditional methods. Their principal function is to provide a visual (or sometimes tactile) feedback channel to assist in the correction of specific speech problems [32]. They are capable of immediately displaying information about a speech utterance, and avoid the difficulties a teacher can have in identifying and verbally describing an error.

Speech training aids can also help alleviate the shortage of highly proficient teachers of speech, by allowing the student to practise with the device on his own. There is even a potential for self-tutoring, as the student may use the device at home. If the device is implemented on a computer and is combined with a program of computer-aided instruction (CAI), then demands on the teacher can be reduced still further. Speech training

aids can also help overcome the problem of adaptation [4], in which the teacher through prolonged exposure to defective speech eventually becomes unaware of its errors.

Recent efforts

Speech training aids go back to the attempts of A.G. Bell in 1874 to use feedback of a deaf pupil's speech waves. The "modern" era began with Bell Telephone Laboratories' visible speech translator of 1944, which was capable of identifying many features of speech. Details of these, as well as recent work on speech training aids, can be found in Pickett [35], Levitt [24], and Pronovost [39]. This section will give only a short survey of some of the more recent devices that have been developed.

In discussing speech training aids, Povel [37] divided them into four categories: pitch and intonation correctors, intensity correctors, rhythm correctors, and articulation correctors. This grouping illustrates both the diversity of problems encountered in the speech of students requiring special training, and the variety of devices that have been proposed.

Numerous pitch displays have been built that give the fundamental frequency of a speech utterance against time. These have proven to be the most useful of the speech training aids, and a number of them are in actual use in deaf schools. Boothroyd reviews some of these in [4], as well as describing one of his own.

Articulation correctors present greater problems to the would-be designer, for no longer is there a single parameter to extract and display. Povel [37] has developed a vowel corrector that helps teach the distinction between the vowels /i/ and /e/. Stark [53] has investigated a method for teaching the production of voiced and unvoiced plosives. Crichton and Fallside [11] have developed an approach for teaching sustained sounds

(especially vowels) that is based on a display of the estimated vocal tract profile for that sound.

Some interesting training aids have come out of Bolt Beranek and Newman, Inc. Nickerson and Stevens [32] have attempted to put together a comprehensive system allowing the display of pitch, intensity, and possibly other parameters, with time. Kalikow and Swets [23] have developed a set of displays for teaching English as a second language (ESL) to Spanish speakers. The displays were carefully chosen in response to common pronunciation errors among Spanish speakers, and show tongue location and trajectory during vowels, vowel duration in multisyllabic words, and amount of aspiration and time lapse before voicing with aspirated-consonant/vowel pairs.

Requirements for speech training aids

The requirements for the successful construction of a speech training aid extend well beyond a straightforward application of speech science and signal processing methods. In addition to the development of basic algorithms for parameter extraction or comparison, it is necessary to design a display modality to present the information in an informative and yet motivating way to the student; to code the algorithms on a mini- or micro-computer to work in real time (assuming it to be possible at all with today's technology); to assemble support hardware, such as a closed-loop tape system for instant replay of a spoken word; to evolve a training program for using the device in a classroom environment; and finally to thoroughly test the effectiveness of the teaching aid, preferably by comparison against a control group of students taught without it.

It is clear that one must work closely with educators and psychologists, and that the work extends beyond "mere engineering". Perhaps

it is the failure of past designers to do this that is responsible for the lack of acceptance of speech training aids by educators [32]. As further evidence that this area of research requires more than an engineering solution stands the experience of Boothroyd [4], who tried unsuccessfully to teach pitch control to a deaf child. He concluded pessimistically that:

> The problems of knowing what to teach, and of structuring the student's environment to create a need for the new skills, may be so great that the exact form which the feedback takes [i.e. the nature of the speech training aid] is of relatively minor importance.

Until the greater problems are solved, the engineer can hope to make no more than a modest contribution to the field.

1.3 OUTLINE OF THESIS

This work comprises an investigation of the feasibility of speech quality analysis via linear predictive analysis. It consists of firstly evolving a suitable definition of speech quality, which ordinarily lacks precise meaning, and then of deriving and testing suitable algorithms for computing quality so defined.

Chapter 2 is concerned with arriving at and justifying a working definition of speech quality. It is necessary to review the speech process and how speech is formed, to examine the well-defined notions that speech pathologists and therapists have of voice and speech quality, and to review the specific speech quality problems of the deaf and of ESL students. A qualitative definition of speech quality is then proposed that characterizes these speech problems in a manner suitable for the design and evaluation of a computer-based speech quality analysis system.

Chapter 3 begins with a short description of the mathematics of linear prediction, the speech analysis method that has been chosen for this work. Linear prediction is excellently suited to digital computation, and is

currently enjoying great popularity in speech processing. Methods of comparing speech utterances are discussed, and a set of distance measures for expressing articulatory and prosodic speech quality is proposed. Finally, the problem of interspeaker differences is examined and a possible method is described for reducing their effect.

Chapter 4 deals with the experimental work that was carried out to evaluate the proposed speech quality measures, and gives results for the performance of the measures under a variety of speech inputs. Experimental reasons for certain choices in the analytic form of the quality measures are also given.

A discussion of the overall significance of the results and their limitations is given in Chapter 5, together with a summary of the findings and directions for further research. Two Appendixes, covering the phonetic alphabet and certain speech processing algorithms, and a bibliography complete this work.

CHAPTER 2

SPEECH QUALITY: MEANING AND MEASUREMENT

2.1 THE SPEECH PROCESS

Speech is the result of using the vocal apparatus to produce sound containing an encoding of linguistically organized thought. Its production involves a complex interaction between mental activity and the dynamic motions of articulatory organs, and has been the subject of much study. An understanding of the means of its production is essential for the appreciation of quality defects in speech. Though this is amply covered in the speech science and linguistics literature (see [15], [10], [12], [54] for discussions and further references), it will be useful to briefly review it here; it will also serve to introduce important terminology.

Speech is produced from the controlled movement of breath from the lungs through the mouth and nose. The breath stream is shaped by the action of the vocal cords (vocal folds), and by the lips, jaw, tongue, and soft palate (velum). Speech comprises the elements of voice, articulation, and prosody. Each of these will be examined in turn.

Voice is the sound produced by the action of the vocal cords on the expiratory breath stream. The vocal cords are folds in the lining membrane of the larynx, and under voluntary control known as phonation, the opening through them (the glottis) can be rapidly opened and closed to produce a quasi-periodic pulsed pressure wave. The lung pressure (subglottal pressure) controls the amplitude of vibration of the vocal cords, and hence the loudness of the resultant sound. The adjustment of length, thickness, and tension applied to the vocal cords determines their fundamental frequency of vibration, and hence the pitch of the sound.

The sound of voice is modified, owing to changes in its harmonic composition, by resonance effects in the vocal tract through which the sound passes. The most important components of the vocal tract are the oral cavity and the nasal cavity. The latter is normally blocked by the velum, but can be coupled to the oral cavity for certain speech sounds, for which the oral cavity is then closed.

The relative positions of the speech organs and resonant cavities are shown in Fig. 2.1.

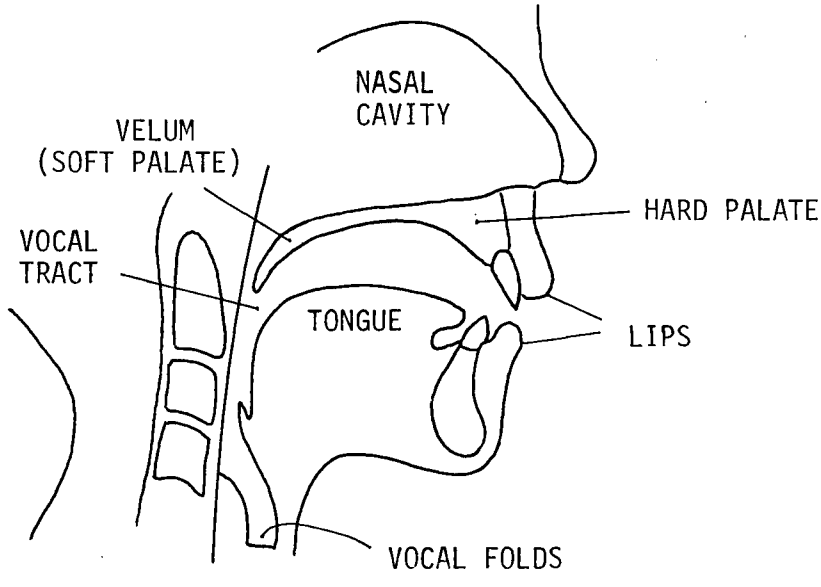


Fig. 2.1 Position of the speech organs
(after Markel & Gray [29])

Articulation is the process of forming from the breath stream the distinct speech sounds of which language is composed. These distinct sounds are called phonemes, and are of two types: vowels and consonants. English uses approximately 45 different phonemes, and these are shown, together with a classification scheme, in Appendix I.

Phonemes are classified according to their manner of production and their place of articulation. The sound may be voiced or unvoiced (i.e. with

or without phonation, respectively), and may be produced by resonance, friction or plosion. The resonants are formed by resonance in the oral (or nasal) cavity, and are all voiced; they include all the vowels, and among the consonants the sonorants (also known as liquids, or as glides and semi-vowels) and the nasals. The remaining consonants form complementary pairs of unvoiced and voiced sounds. The fricatives are formed by rapidly forcing air through a small constriction so as to give turbulent, noisy, flow ('audible friction'). The plosives are generated by the sudden release of built-up air pressure behind an occlusion in the vocal tract. Because of the motion required to form them, they are the only simple sounds not capable of being sustained. A combination sound known as an affricate is formed when the release of air pressure is relatively slow and audible friction occurs.

The vowels are controlled mainly by tongue position, and all being resonants are classified according to tongue hump position and tongue height or degree of restriction. Other factors affecting the vowels are lip rounding, whether the tongue muscles are tense or lax, etc. The consonants' second dimension for classification is their principal place of articulation, which for the sounds of English can be: the two lips (bilabial), the upper teeth on the lower lip (labiodental), the tongue behind the teeth (dental), the tongue to the gum ridge (alveolar), the tongue against the hard palate (palatal) or the soft palate (velar), and the vocal cords constricted and fixed (glottal).

Because of the dynamic constraints imposed on the movement of the articulatory organs, particularly the tongue, phonemes are influenced by adjoining ones. For example, the /k/ in 'kid' is distinctly different from the one in 'could'. The phenomenon is known as coarticulation, and is useful in the perception of speech because of the information it gives about the adjoining sounds. Different forms of the same phoneme are known as

allophones.

Prosody is the term used to refer to the rhythm, stress, and intonation components of speech that are used both for communicating additional linguistic information and for conforming to established conventions of language. For example, syllable stress is required in every English word of more than one syllable, and phrasing is used in much the same way as is punctuation in written language.

The acoustical correlates of prosody (sometimes referred to as the lower level prosody [56]) are vocal pitch, intensity, and phonetic duration. Syllable and word stress are accomplished by a rise in pitch and an increase in vowel duration, together with an increase in intensity. Intonation involves a rise or fall in pitch towards the end of a sentence, and is best characterized by the pitch contour of the sentence. Phrasing is the insertion of silent intervals ('boundaries') into a continuous utterance.

2.2 COMPONENTS OF SPEECH QUALITY

The term "speech quality" is used in this work to mean, loosely, the degree of perfection present in the speech being appraised. Although electrical engineers have long had to compare speech transmission systems on the basis of which sounds best, and speech clinicians are directly concerned with the diagnosis and treatment of speech disorders, neither these groups nor others have satisfactorily defined speech quality. However, an understanding of what is speech quality can be gained by examining their differing approaches to the question.

Electrical engineers have a measurement approach to speech quality, involving the use of preference and intelligibility tests, e.g. [21], [16]. The dimension of preference is essentially that of pleasantness or

mellisonance; the term aesthetic acceptability has also been used. Generally, defects of voice affect its mellisonance only and not its intelligibility (though they may have a distracting effect). Defects of articulation or prosody can greatly affect the intelligibilty of the speech as well as its mellisonance.

Mellisonance is assessed via preference tests, in which speech samples are ranked in order of preference from a series of two-way comparisons, or by category judgment, in which each speech sample is rated (e.g. from unsatisfactory 0% to excellent 100%) according to an arbitrarily assigned scale. Intelligibility is measured as the fraction of words understood correctly in test phrases. Because of the contextual cues present in continuous text, isolated words or short phrases must be used.

In contrast, speech clinicians have viewed speech quality from an analytical viewpoint. The defects of speech are the defects in its component elements of voice, articulation, and prosody [5]. The remainder of this section examines these factors in more detail.

Defects of voice

The dimensions of voice are pitch, loudness, vocal quality, and nasality, and hence defects of voice involve problems with one of these. Both pitch and loudness are prosodic variables (and therefore are linguistically important, unlike voice), but their average values and their range of variation are attributes of voice only. Vocal quality, also called voice quality, is a term universally used in speech pathology to describe the timbre or tone of the voice, typically being expressed via words such as harsh, hoarse, strident, resonant, etc. It includes the effects of vocal mode (part-fold, as in vocal fry and falsetto, to full-fold or normal) and vocal constriction (from open to closed). The term voice, as usually used,

includes only those aspects of the speech production process that are not phonemically significant. Thus hypernasality is a problem of voice as it merely results in speech having an unpleasant nasal ring, whereas hyponasality, in which the inadequately nasalized phonemes /m/, /n/, /ŋ/ can be confused with the non-nasalized phonemes /b/, /d/, /g/, is properly a defect of articulation.

Defects of articulation

These are due to errors with voicing, manner of production, and place of articulation, and errors or inaccuracies with the dynamics of forming individual sounds and combinations of sounds. A great variety of defects have been reported in the literature, and a representative selection follows; actual errors among the deaf and among ESL students are discussed in section 2.3. Errors of voicing involve the substitition of a voiced phoneme for an unvoiced one, e.g. /d/ for /t/, and vice versa. Errors of manner of production include hyponasality and such effects as replacement of fricatives by plosives or affricates, e.g. /t/ for /θ/. Errors of place of articulation are more common, including distortions of recognizable phonemes, substitutions of similar sounds (e.g. /w/ for /r/, /ə/ for /e/, etc.). Errors of dynamics include diphthongization of pure vowels, malarticulation of consonant blends, inaccuracies with voice onset after unvoiced consonants and with the timing of general transitions between sounds, etc.

Many other schemes for classifying articulatory disorders are possible [38]. A simple one groups them as omissions, substitutions, distortions, and additions, where substitutions and distortions are distinguished according to whether or not the sound produced is phonemic.

Defects of prosody

These are errors in syllable and word stress, intonation, and rhythm, such as monotone pitch, irregular or erratic stress and rhythm, etc. The importance of correct prosody has been demonstrated by Hudgins and Numbers [20] in their investigation of the speech of the deaf: a sentence spoken with correct stess and rhythm was almost four times as likely to be understood as one without.

2.3 SOME SPECIFIC SPEECH QUALITY PROBLEMS

This section will give a brief review of the specific speech problems that have been noted in investigations of the speech of the deaf, and of some of the pronunciation difficulties faced by students of English as a second language.

(1) Speech quality problems among the deaf

An important distinction can be drawn between the prelingually deaf - those who were born deaf or who lost their hearing prior to the development of speech (at around age 3) - and the postlingually deaf, who suffered their hearing loss in later life. By deafness here is meant a hearing impairment of about 90 dB (threshold level), which is sufficient to render "everyday auditory communication impossible or nearly so" [12]. The greatest difficulties with speech occur with prelingual deafness, and the object of speech training is the development of speech. With deafness in later life, training is aimed at the preservation of speech.

The defects in the speech of (prelingually) deaf children have been investigated by many researchers, both by comparison with the speech of normally hearing children, and independently by correlating defects with

speech intelligibility. Comprehensive studies include those by Hudgins (1934) [19], Hudgins and Numbers (1942) [20], and Calvert (1961) [6]. Summaries of reported problems can be found in [52] and [32].

Most characteristic of the speech of the deaf is their "distinctive voice quality" [7]. It has regularly been described as tense, flat, breathy, and throaty, and it has even been suggested as a clinical indicator of deafness. In [19], Hudgins reported the speech of the deaf to be characterized by slow and laboured speech with extensive expenditure of breath, resulting in short, irregular breath groups. Vowels and fricatives, indeed entire sentences, are prolonged to 2 to 4 times their normal length, and there is excessive nasality with both consonants and vowels.

In their investigation of the intelligibility of the speech of deaf children, Hudgins and Numbers [20] found both articulatory and prosodic errors to be responsible for poor intelligibility. Errors of articulation involving the consonants were voicing errors, consonant substitutions, malarticulation of compound and of abutting consonants, and omission of arresting and of releasing consonants. The most difficult consonants to pronounce correctly were (in order of difficulty) /dʒ/, /d/, /h/, /b/, /g/, /ʃ/. Among the vowels the problems were vowel substitutions, malarticulation of diphthongs, and diphthongization or neutralization of vowels. The difficult vowels were /aI/, /ɔI/, /ɜ/, /i/, /ɛ/. Errors of prosody included misplacement or absence of word and syllable stress and of phrase-level boundaries, incorrect intonation patterns, and the inability to control pitch and loudness independently. Utterances frequently lacked a natural rhythm.

Postlingual hearing loss can also cause serious defects in speech quality. Articulatory defects generally occur first, typically involving distortion of unvoiced fricatives and omission of arresting consonants [51].

Abnormalities of voice quality and with use of prosody can follow. However, deterioration of speech quality can be minimized by a program of speech conservation, which usually takes the form of developing the subject's sensitivity to the kinesthetic cues accompanying speech, a program which in some respects is not unlike that used to teach the prelingually deaf child.

(2) Speech quality problems with ESL

A foreign accent can be considered a speech quality defect. The speaker is unable to make his pronunciation and rhythm conform to the requirements of the second language as a result of his enormous familiarity with his native language. Whereas, with the deaf, speech quality problems arise from an inability to hear their own and others' attempts at speech, the problem for ESL learners is interference from the sound system and rules of their native language [34], [36].

The pronunciation (articulation) problems may be classified as follows.

(1) The phoneme does not occur in the speaker's native language. Examples of this abound: the vowels /æ/, /I/ are absent in many languages, as are the consonants /θ/, /ð/, /ʍ/, /ʒ/, /dʒ/; the French and Spanish lack /h/, the Germans lack /w/, the Japanese and Chinese lack both /r/ and /l/; Spanish is also missing /U/, /ɔ/, /ə/, /ʃ/, /v/. In this case, the speaker substitutes the nearest familiar sound or its orthographical equivalent in his native language (e.g. /v/ for /w/ with Germans).

(2) The phoneme is articulated differently in English. For example, in French and Spanish /t/ is dental, but in English it is alveolar. Moreover, initial /t/ is not aspirated, whereas in English it is; an unaspirated initial /t/ sounds very much like a /d/ to an English listener. The English /r/ is very different in character to that of

other languages. Here the speaker substitutes his native articulation.

(3) The utilization of the phoneme is different in English. Thus, in Spanish /z/ occurs only as an allophone of /s/, used before a voiced consonant, and /ð/ occurs only as an allophone of /d/ used between vowels. German has no voiced consonants at the ends of syllables, using instead the unvoiced counterpart. Consonant clusters, e.g. /mpst/ in glimpsed, are common in English, but do not exist in Japanese, or in Spanish (in word-final position), and are another great source of difficulty.

Problems with prosody arise from differences in the use of stress, pitch and juncture. In French, each word and each word group is given an increase in stress towards its end, and if this is carried over to English, the result sounds poor. Spanish lacks both the diphthongization of stressed vowels and the neutralization of unstressed vowels, a feature of English. Juncture is rare in both Spanish and French.

2.4 PRACTICAL CONSIDERATIONS

The preceding discussion has shown speech quality - the degree of perfection present in the speech - to be the lack of defects in its components of voice, articulation, and prosody. This is a qualitative definition based on a comparison with "normal" speech, but one that is quite adequate for the purposes of speech training, where the requirements are essentially diagnostic and corrective. A single-valued quantitative measure of speech quality, useful as it may be in Communications Engineering, is unable to satisfy these requirements. Moreover, it can only be obtained after an investigation of how speech quality defects affect mellisonance and intelligibility scores, a separate piece of research that, though useful, is not needed for speech quality analysis.

To illustrate, a speech quality system for speech training need not judge, for example, the relative severity of a consonant distortion as against misplaced word stress, but should be capable of distinguishing misplaced stress from inadequate or absent stress.

As a first step in investigating the feasibility of speech quality analysis, it will be acceptable to ignore defects of voice, and instead to concentrate on developing techniques sensitive to articulation and prosody errors. Poor voice quality is more difficult to correct without the skilled interaction of a speech clinician, and fortunately is not an important contributor to poor speech intelligibility.

There are other concessions that need to be made to practice. The first of these concerns the way in which quality is assessed. I propose to view speech quality as inherently relative, and to judge it only by direct comparison between the test utterance and one produced by a "teacher" speaking the same text. This approach carries with it the disadvantage that any features of speech unique to the teacher (such as a regional accent) are taken as standard, their absence in the test utterance being flagged as a quality defect. Also, in a practical situation, some flexibility will be lost as a teacher must be present or must have produced a tape of speech exercises. But the advantages of simplicity and precision afforded by having a direct standard available appear to outweigh the disadvantages.

Another concession is the restriction of analysis to single words and short phrases only. The present expense and amount of real time used to process the vast amount of data in speech precludes the analysis of full sentences. This severely restricts the extent to which sentence level prosody can be assessed and corrected. However, much useful speech training can be done at the word and phrase level, and in any event, the computing restriction is likely to be only a temporary one.

The speech processing method of linear prediction, to be described in Chapter 3, appears to be well suited to the analysis of speech articulation and prosody. Linear prediction reduces the data to manageable size, and nicely reflects its spectral properties (including implicitly therewith the shape or position of the articulators, and the identity of the speech sounds actually made). LP techniques also allow the loudness and pitch of speech, whether or not it is voiced, and its time alignment with respect to a reference utterance to be monitored. Additionally, the fields of speech recognition and speaker identification have contributed numerous techniques for comparing two samples of LP-processed speech.

Yet the requirements for speech quality analysis are different from those for speech or speaker recognition, and the suitability of these techniques has yet to be investigated. I propose to examine in subsequent chapters the following three questions. Satisfactory answers to each will go a long way towards establishing the feasibility of automatic speech quality analysis as envisioned above.

(1) Can linear prediction analysis of speech reliably detect articulation errors? If so, what kind of errors, and using which techniques?

(2) Do interspeaker differences mask these quality differences, and if so, how might their effect be reduced?

(3) Can a linear prediction analysis of speech reliably detect prosody errors, especially general timing errors?

CHAPTER 3

LINEAR PREDICTION AND SPEECH QUALITY

3.1 LINEAR PREDICTION OF SPEECH

A digital model of speech production

The speech production mechanism can be modelled as an acoustical tube of varying dimensions (the vocal tract) that is excited at one end by a glottal pulse or noise source, and terminated at the other by the lips. The acoustical tube acts as a linear time-varying filter, and for convenience is assumed to include the spectral effects of glottal flow and lip radiation. The model is essentially due to Fant [14], and is depicted in Fig. 3.1.

The glottal source consists of periodic pulses when the sound is voiced, and of random noise when it is not. Despite some deficiencies, such as the need to regard nasal sounds as arising from excitation of the vocal tract, the model has proved very successful in a wide variety of applications, synthesis as well as analysis. In the linear prediction model,

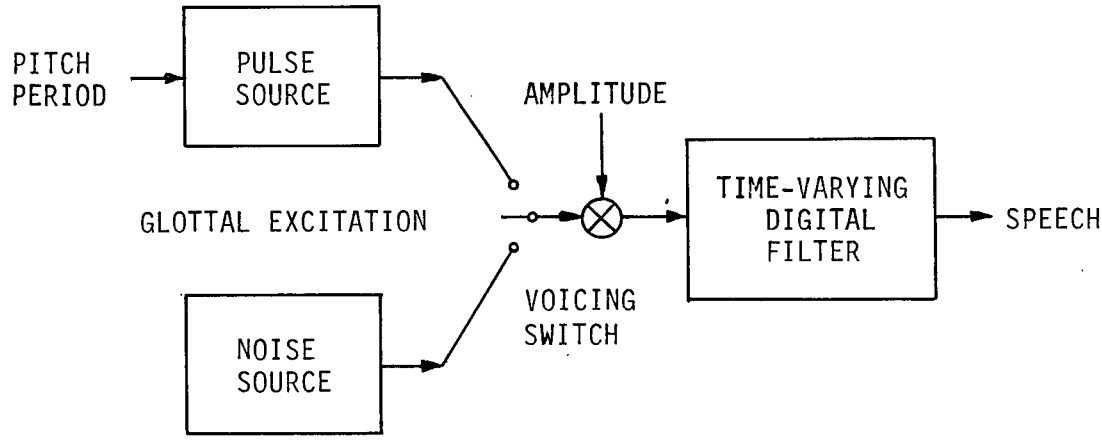


Fig. 3.1 Digital model of speech production (after Schafer & Rabiner [49])

made popular by Atal and Hanauer [3] in 1971, the filter is represented by its z-transform, and its coefficients are estimated by linear prediction on the sampled values of actual speech. The filter is assumed to be stationary over short periods of time, typically 10 ms to 30 ms; its order is generally chosen to lie between 8 and 14.

Calculating the filter coefficients

In the model of Fig. 3.1, let the excitation signal be u(n), n=0(1)N-1, the amplitude be σ, and the filter coefficients be $\{a_k\}$, k=0(1)p with a_0=1. p is the filter order. The speech signal is then given by

$$X(z) = \frac{\sigma}{1 + \sum_{k=1}^{p} a_k z^{-k}} \, U(z) \tag{3.1}$$

or

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + \sigma u(n) \tag{3.2}$$

The $\{a_k\}$ are found by minimizing the total squared error E that arises from predicting x(n) from a linear combination of past values only.

$$E = \sum_{n} e^2(n) \tag{3.3}$$

where

$$e(n) = \sum_{k=0}^{p} a_k x(n-k) \tag{3.4}$$

Therefore

$$E = \sum_{i=0}^{p} \sum_{k=0}^{p} a_i a_k \sum_{n} x(n-i)x(n-k) \tag{3.5}$$

Two choices of summation over n are possible, giving rise to two different solutions for the $\{a_k\}$ [25]:

(a) autocorrelation method (Yule-Walker method): AC

Here we choose $-\infty<n<\infty$, and assume x(n)=0 for n<0 and n>N-1. In practice this assumption is met by the use of a finite duration window w(n) which premultiplies x(n). Setting $\partial E/\partial a_k = 0$ in Eq. (3.5) gives a set of

linear equations in the a_k:

$$\sum_{k=1}^{p} a_k r_{|i-k|} = - r_i, \qquad i=1(1)p \tag{3.6}$$

where $\{r_i\}$ is the autocorrelation sequence of $x(n)$:

$$r_i = \sum_{n=0}^{N-1-i} x(n)x(n+i), \qquad i=0(1)p \tag{3.7}$$

r_0 is the energy in the speech frame, and is proportional to σ^2.

Eq. (3.6) can be solved by Gaussian elimination, but there is a more efficient procedure known as the Levinson method that makes use of the special form of the equations. It is described in Appendix II.

With the $\{a_k\}$ satisfying Eq. (3.6), the total squared error E is minimized, and is given by

$$\propto = E_{min} = r_0 + \sum_{k=1}^{p} a_k r_k \tag{3.8}$$

$\propto$ is known as the prediction residual, and represents the residual energy in the output of the inverse filter $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ operating on the speech signal $x(n)$.

Prior to calculation of the autocorrelation coefficents, it is usual to pre-emphasize $x(n)$ by differencing once in the time domain (i.e. multiplying $X(z)$ by $1-z^{-1}$ in the frequency domain). This cancels one of the poles due to glottal flow, and experimentally has been found to give improved results in most applications. $x(n)$ is additionally preprocessed by windowing, so as to taper the data smoothly to zero at the ends of the finite sample. The window most commonly used is the Hamming window, given by

$$w(n) = 0.54 - 0.46 \cos 2\pi n/(N-1), \qquad n=0(1)N-1 \tag{3.9}$$

(b) <u>covariance method</u> (least squares method): COV

Here we choose $p<n<N-1$, and in consequence $x(n)$ is always known and

windowing is unnecessary (though x(n) is still pre-emphasized). The resulting equations are again linear, and are:

$$\sum_{k=1}^{p} a_k \phi_{ik} = - \phi_{i0}, \qquad i=1(1)p \tag{3.10}$$

where the ϕ_{ik} are the covariances of the x(n):

$$\phi_{ik} = \phi_{ki} = \sum_{n=p}^{N-1} x(n-i)\,x(n-k), \qquad i,k=0(1)p \tag{3.11}$$

Eq. (3.10) cannot be solved by the Levinson method, but as the resulting matrix of coefficients is symmetric and positive definite, Cholesky decomposition can be used to gain some improvement in computational efficiency over Gaussian elimination.

The prediction residual is given by

$$\alpha = E_{min} = \phi_{00} + \sum_{k=1}^{p} a_k \phi_{k0} \tag{3.12}$$

The covariance (COV) method, by avoiding the need for windowing, is able to estimate the filter coefficients from the data with considerably greater accuracy than the autocorrelation (AC) method; de Souza [13] has presented results that amply confirm this. However, the COV method can sometimes give rise to an unstable filter, while the AC method, for sufficient numerical accuracy, will always give a stable filter [29]. Choice of which method to use in a particular application is therefore based on which property is deemed more important. Computational efficiency is not really a factor in the decision, for the methods, surprisingly, require similar amounts of computation. The dominant aspect of both methods, for $N \gg p$, is calculation of the covariances or autocorrelations, and Appendix II shows that the matrix of covariances can be computed in almost the same time as the array of autocorrelations.

Transformations of the filter coefficients

There are a number of important transformations of the filter coefficients that can uniquely characterize the linear prediction filter $H(z) = X(z)/U(z)$. These alternative parameter sets are related to the filter coefficients by a 1:1 non-linear transformation, and have distinct physical interpretations. The most useful ones for speech processing are:

1. The filter coefficients $\{a_k\}$, k=1(1)p.

2. The normalized autocorrelation coefficients $\{r_i\}$, i=1(1)p with $r_0=1$, of the impulse response of the filter. Conversion from $\{a_k\}$ to $\{r_i\}$ is described in Appendix II.

3. The reflection or "parcor" (partial correlation) coefficients $\{k_i\}$, i=1(1)p, defined by $k_i=a_i^{(i)}$ where $a_i^{(i)}$ is the i th filter coefficient of an i th order filter fitted to the speech data. k_i can be considered the reflection coefficient at the boundary between sections i and i+1 of a p-section acoustic tube having transfer function H(z).

4. The log area coefficients $\{g_i\}$, i=1(1)p, defined as

$$g_i = \log (1+k_i)/(1-k_i) \tag{3.13}$$

Note that $g_i = \log(A_i/A_{i+1})$ with $A_{p+1}=1$, where A_i is the cross-sectional area of the i th section of the acoustic tube model.

5. The poles $\{z_i\}$, i=1(1)p, of the filter H(z), defined by

$$\prod_{k=1}^{p} (1-z_k z^{-1}) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.14}$$

6. The cepstral coefficients $\{c_i\}$, i=1(1)p, of H(z), defined by [1],[33]

$$\sum_{k=1}^{\infty} c_k z^{-k} = - \ln (1 + \sum_{k=1}^{p} a_k z^{-k}) \tag{3.15}$$

giving
$$c_1 = - a_1 \tag{3.16}$$
$$c_k = - a_k - \sum_{i=1}^{k-1} (1-i/k) c_{k-i} a_i, \qquad k=2(1)p$$
$$= - \sum_{i=1}^{p} (1-i/k) c_{k-i} a_i, \qquad k=p+1,\ldots$$

Each of these parameter sets represents a different weighting of the properties of the speech signal from which it is derived, so certain applications will favour certain parameter sets.

Calculating loudness, pitch, and other speech properties

The LP parameters characterize the spectrum of the speech during the speech frame. To fully describe the speech, it is necessary to also specify the loudness, pitch, and voiced-unvoiced nature of the speech over the frame;[1] these aspects of speech can be determined by LP-related methods. Also, it is of interest to obtain from the LP parameters the traditional descriptions of the speech spectrum; these can be determined directly. This section discusses how these speech properties are obtained.

The loudness L of speech is simply equal to the energy r_0 or ϕ_{00} in the speech frame. It is usual to express L in decibels relative to some reference level R_0, so we take

$$L = 10 \log_{10} r_0/R_0 \tag{3.17}$$

where
$$r_0 = \sum_{n=0}^{N-1} x(n)^2 \tag{3.18}$$

---

[1] Loudness and pitch in this work are synonyms for the average energy and fundamental frequency of speech; they are not the subjective quantities of the same name used in acoustics and audiology. Note that pitch is undefined for unvoiced speech.

The pitch P of speech is defined here to be

$$P = \log_2 f_0/F_0 \tag{3.19}$$

where f_0 is the fundamental frequency of the vocal tract excitation. This expresses it in octaves above some reference pitch F_0. Determining the pitch of speech essentially relies in the observation that the residual e(n), defined in Eq. (3.4), is equal to the excitation function u(n), and should show a pronounced peak every pitch period. This allows one to find the pitch period, or in the event of no regular peaks in e(n), to deduce unvoiced speech. The actual problems in deriving pitch period or voicing function, however, are considerable, and special algorithms are required to obtain reliable estimates. These are discussed by Markel in [27] and [29,Ch.7], and Rabiner in [43]. Important as knowledge of pitch is to calculations with speech prosody, the implementation of a pitch extractor was felt to be unwarranted at this stage of the work.

Traditional represention of speech include its short term spectrum and a formant analysis. The frequency response of speech is readily found from

$$H_S(j\omega) = \frac{\sigma}{1 + \sum_{k=1}^{p} a_k e^{-j\omega Tk}} \tag{3.20}$$

Formants are the resonant frequencies of the vocal tract, and the first 3 to 5 formants, with their bandwidths, are normally sufficient to characterize articulation. In general, the formants correspond to the poles z_i of H(z), defined in Eq. (3.14). However, sometimes it is difficult to associate a particular formant with a pole; a discussion of the difficulties and their solutions can be found in [26] or [29,Ch.7].

3.2 METHODS OF COMPARING SPEECH UTTERANCES

The previous section has described LP methods for characterizing speech over a single time frame. This section considers methods for comparing two complete speech utterances. From this comparison will arise techniques for evaluating the speech quality of one utterance with respect to the other.

One speech segment will be taken as the test utterance and will be described by umprimed symbols; the other, the reference utterance, will be described by primed symbols. If the vector of LP coefficients for a single time frame m of the test signal $\underline{S}$ is denoted by $\underline{a}(m)$, then

$$\underline{S} = \{\underline{a}(m) \mid m=1(1)M\} \tag{3.21}$$

where M is the total number of (possibly overlapping) frames in the test utterance. Similarly, the reference signal $\underline{S}'$ is given by

$$\underline{S}' = \{\underline{a}'(n) \mid n=1(1)N\} \tag{3.22}$$

with n, N written in place of m', M'

We are seeking functions $f_i(\underline{S},\underline{S}')$ for comparing the two utterances that describe in some meaningful way the quality of $\underline{S}$ with repect to $\underline{S}'$. It is clear from the discussion in Chapter 2 that the f_i fit into two distinct classes: measures of articulatory quality, and measures of prosodic quality. Since quality is defined non-quantitatively, there will not be a strict correlation between measures of quality and actual speech quality; nevertheless it will be clear that the f_i do measure the dissimilarities that influence subjective assessments of quality.

The f_i will be functions of the results of comparing individual frames of $\underline{S}$ and $\underline{S}'$. Thus

$$f_i(\underline{S},\underline{S}') = f_i(\{d_i(w(n),n) \mid n=1(1)N\}) \tag{3.23}$$

where $d_i(m,n)$ is a scalar measure of the dissimilarity between test frame m and reference frame n, and $w(n)$, $n=1(1)N$, is a mapping function that maps the reference time axis into the test time axis. The function d is referred to as a distance measure, and the function w as a time warping or time normalization function. The remainder of this section will be concerned with arriving at suitable choices for the $d_i(m,n)$, $w(n)$, and f_i.

Distance measures for articulatory quality

Articulation is almost completely characterized by the spectrum of the speech.[2] As the filter coefficients are sufficient to fully specify the spectrum of (non-nasalized) speech, it would appear that distance measures based solely on the $\{a_k\}$ (or a transformation of them) would be adequate as indicators of articulatory quality. Thus we may take

$$d_A(m,n) = d_A(\{a_k\},\{a_k'\})$$

Distance measures of this kind have been used very frequently for speech recognition (cf. [2], [18], [40]), and four of the more successful or promising of these are reviewed next.

(1) Prediction residual ratio RESID

This is a simple function of the filter and autocorrelation coefficients proposed by Itakura [22] in 1975 and used extensively since. The distance measure is taken to be the natural logarithm of the ratio of the prediction residual δ obtained by passing the test signal through the inverse filter $A'(z)$ to the minimum residual α obtained by passing the test signal through its inverse filter $A(z)$. Thus

---

[2] Voicing function, i.e. whether or not the speech is voiced, is also required.

$$d_A(m,n) = \ln \delta/\alpha \tag{3.24}$$

where

$$\alpha = r_0 + \sum_{k=1}^{p} a_k r_k \tag{3.25}$$

$$\delta = \sum_{i=0}^{p} \sum_{k=0}^{p} a_i' a_k' r_{|i-k|} \tag{3.26}$$

δ/α is the ratio of a prediction residual to a minimum prediction residual, and is a likelihood ratio under certain circumstances.

(2) Cosh measure COSH

The Itakura measure has received much criticism from theoreticians and statisticians, e.g. [13], [17], because it is asymmetric, i.e. a different distance is obtained if the test and reference frames are interchanged. Gray and Markel [17] have combined the two likelihood ratios δ/α and δ'/α' to obtain a theoretically sound symmetrical measure. The cosh measure is

$$d_A(m,n) = \cosh^{-1}(\delta/\alpha + \delta'/\alpha')/2 \tag{3.27}$$

where α and δ are as before, and

$$\alpha' = r_0' + \sum_{k=1}^{p} a_k' r_k' \tag{3.28}$$

$$\delta' = \sum_{i=0}^{p} \sum_{k=0}^{p} a_i a_k r'_{|i-k|} \tag{3.29}$$

$$\cosh^{-1} x = \ln(x + \sqrt{x^2-1})$$

The cosh measure has the property that it approximates, and bounds from above, the rms difference between the log spectra of the two speech signals. It therefore provides a highly efficient technique for calculating that difference. (To express it in decibels, the cosh measure must be multiplied by $10/\ln 10 = 4.34$.)

(3) Cepstral measure CEPS

The cepstral measure, also proposed in [17], is an alternative method of calculating the rms difference between the log spectra, bounding it from below. The distance measure is taken as

$$d_A(m,n) = \sqrt{2 \sum_{k=1}^{2p} (c_k - c_k')^2} \tag{3.30}$$

where the cepstral coefficients $\{c_k\}$ are as defined in Eq. (3.16). Taking the infinite rather than partial sum in Eq. (3.30) gives the rms spectral measure exactly (in dB if multiplied by 4.34).

(4) F-test measure FTEST

De Souza [13] has derived a statistical test for comparing two segments of speech from their LPC coefficients (the filter coefficients obtained by the COV method). The test computes a statistic F of known distribution that can be used to test the null hypothesis that the two observed series (speech samples) arise from the same process. The distance measure is therefore taken to be

$$d_A(m,n) = - \ln Q_F(F \mid p, 2N-4p) \tag{3.31}$$

where

$$F = ((2N-4p)/p) \ (\bar{\alpha}/(\alpha+\alpha')-1) \tag{3.32}$$

and Q_F is the upper-tail-area function for the F distribution with p and 2N-4p degrees of freedom. N is the number of speech samples in a speech frame, α and α' are the prediction residuals for the test and reference signals, and $\bar{\alpha}$ is the prediction residual for the two signals combined together into one series:

$$\bar{\alpha} = \bar{\phi}_{00} + \sum_{k=1}^{p} \bar{a}_k \bar{\phi}_{k0} \tag{3.33}$$

where $\bar{\phi}_{ik} = \phi_{ik} + \phi'_{ik}$, and $\bar{a}_k$ satisfies $\sum_{k=1}^{p} \bar{\phi}_{ik} \bar{a}_k = - \bar{\phi}_{i0}$ for i=1(1)p.

The time warping function w(n)

The function w(n), which maps the reference time axis [1,N] into the test time axis [1,M] determines which frames of the test and reference utterances to compare. If M=N, then the simplest possible choice for w is w(n)=n. In general, a linear map from reference to time axis is

$$w(n) = 1 + [(n-1)(M-1)/(N-1)], \quad n=1(1)N \tag{3.34}$$

But a linear mapping function cannot take account of any variation in speaking rate that may exist between the two utterances. Particularly with multisyllabic words, there may be imperfect registration in time between the phonemes. For example, vowel duration can be incorrect because of wrong or misplaced syllable stress, or because of neutralization or diphthongization of the vowel. The actual time alignment pattern is therefore an important indicator of prosodic quality, and it is desirable to choose w(n) to approximate it, e.g. Fig. 3.2.

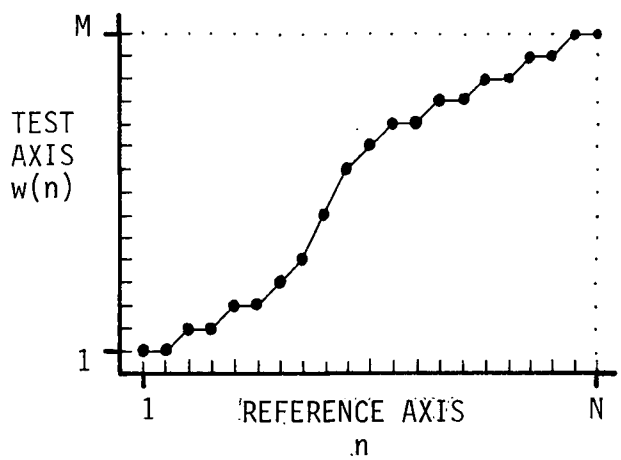


Fig. 3.2 A non-linear time warping function

Therefore, we choose w(n) to optimize the agreement between the two utterances, as expressed via the function

$$D = \sum_{n=1}^{N} d_A(w(n),n) \tag{3.35}$$

subject to the endpoint and continuity constraints

1. $w(1)=1$, $w(N)=M$ (3.36)
2. $w(n+1)-w(n) = 0$, 1, or 2 if $w(n)\neq w(n-1)$, else 1 or 2.

The optimum w(n) can be found by dynamic programming on Eq. (3.35). An algorithm for this is listed in Appendix II. The idea of dynamic programming (DP) and non-linear mapping functions to compensate for imperfect time alignment between utterances is due to Sakoe and Chiba [46] who used them to obtain improved performance in a speech recognition system. Their use has since become commonplace in speech recognition systems.

Distance measures for prosodic quality

Prosodic quality is a function of the loudness, pitch, and timing of one utterance with respect to another. As the average value and range of variation of these quantities are attributes of voice and not prosody, it is necessary to include factors in the distance measures for prosody to cancel out and equalize them. These factors must be estimated from data obtained across the entire utterance, so unlike distance measures for articulation, the ones for prosody are utterance dependent as well as frame dependent.

(1) Loudness

If we restrict ourselves to factoring out differences in average value and range of variation via a linear transformation of the loudness in decibels, then a suitable loudness distance measure is

$$d_L(m,n) = a_1 L(m) + a_2 - L'(n) \tag{3.37}$$

The transformation variables a_1 and a_2 are chosen to give a minimum rms value of d_L across the utterance, i.e. a_1 and a_2 minimize

$$D_L = \sum_{n=1}^{N} (a_1 L(w(n)) + a_2 - L'(n))^2$$

This results in the values

$$a_1 = (N\sum LL' - \sum L \sum L')/(N\sum L^2 - (\sum L)^2) \tag{3.38}$$
$$a_2 = (\sum L^2 \sum L' - \sum L \sum LL')/(N\sum L^2 - (\sum L)^2)$$

(2) Pitch

In analogous fashion, define a pitch distance measure to be

$$d_P(m,n) = a_3 P(m) + a_4 - P'(n) \tag{3.39}$$

where a_3 and a_4 are defined similarly to a_1, a_2 in Eq. (3.38).

(3) Timing

If w(n) is chosen optimally, then its derivative w'(n) will be a good indicator of the instantaneous speed of the test utterance relative to the reference utterance. Values of w'(n)>1 mean that the test word is being spoken more slowly than the reference word, and w'(n)<1 that it is being spoken more quickly.

The quantity (M-1)/(N-1) represents average speed of the test to the reference utterance, so we take the timing distance measure to be

$$d_T(m,n) = w'(n) - (M-1)/(N-1) \tag{3.40}$$

Because w(n) is an integer-valued function defined over the integers, w'(n) cannot be obtained by normal differentiation. A simple numerical differentiation formula such as w'(n)=w(n)-w(n-1) can be shown to greatly amplify the round-off noise present in w(n) that comes from representing a

continuous relationship by an integerized function. The ideal differentiation formula smoothes out such local irregularities in w(n), but responds rapidly to more extensive changes in its behaviour. Experimentation is needed to find the most suitable such formula, and results for this are reported in Section 4.3.

Combining individual distance measures

The individual distance measures d_i, evaluated at each pair of frames (w(n),n), need to be combined to form overall quality measures f_i. Each f_i describes a particular aspect of the speech quality between the test utterance S and the reference utterance S'. The f_i are the outputs of the speech quality analysis system, and must therefore convey to the user all the available and desired information about the quality evaluation. It follows then that selection of the f_i depends on sufficient knowledge being available as to (i) the kind of outputs desired of a speech quality analysis system, and (ii) the relationship of the observed distance measures across a word pair to the actual errors they represent. For example, some applications may require the full set of distance measures $d_i(n) = d_i(w(n),n)$ for n=1(1)N, while others require only a simple judgment of good or bad quality; in some contexts a large value of d_A occurring for only a few frames may have special meaning, while in other contexts it may be without significance. It is therefore important to defer decision as to the form of the overall speech quality measures f_i until all the required knowledge has been derived.

A simple but adequate choice of f_i for representing the results of the distance measure evaluations of Chapter 4 is to take each f_i to be the set $\{d_i(n) \mid n=1(1)N\}$, evaluated against a fixed threshold t. The length of the period for which $d_i(n) > t$ indicates approximately the duration of the

quality error, and the magnitude of $d_i(n)$ in the interval indicates its severity.

3.3 INTERSPEAKER DIFFERENCES IN SPEECH

Two segments of speech spoken by different persons, and judged subjectively to be of good quality, will show substantial differences when compared through the techniques described above. These differences are known as interspeaker differences, and they arise from dissimilarities in the physiology of the vocal apparatus and in learned patterns of movement of the articulators; for example, a shorter vocal tract length will result in higher formant frequencies. Interspeaker differences are minimized by the removal of average value and range of variation in the prosodic distance measures. However, the LP parameters on which the articulatory distance measures are based are quite speaker dependent. They have even been used successfully for speaker identification [1]. For this reason one can expect that the articulatory distance measures described in this chapter will be speaker dependent.

Speaker dependence in speech comparisons is also very much a concern in the field of speech recognition. Most of the systems implemented to date are in fact single speaker systems - the speaker who wishes to use the system must be the one to speak the reference vocabulary. But some speaker-independent speech recognition (SISR) systems have been built, and it can be expected that a study of the emerging methods used to overcome interspeaker differences in SISR will reveal techniques applicable to speech quality analysis.

Unfortunately this is not the case. The method of multiple reference templates per word, as described by Rabiner in [40] and Gupta in [18], is not practicable because in speech quality analysis there is only one teacher

to speak the reference utterance. Moreover, sensitivity to quality differences is lost by simply matching the test word to the nearest reference word. The method used by Sambur and Rabiner in [48] to achieve SISR for spoken digits is also not applicable, as it bases its classification on crude speaker-independent measurements (e.g. "five" starts with a fricative and "eight" with a vowel) that cannot capture more subtle quality differences.

It therefore appears that new techniques are required for cancelling interspeaker differences in speech quality analysis. A thorough investigation of possible techniques, however, is beyond the scope of this work, for basic questions remain to be answered concerning the performance of the regular articulatory measures and the effect on them of differences between speakers. One method that does deserve investigation now is that of orthogonal linear prediction, which appears to offer a means of reducing speaker dependence in the LP parameters themselves. It comes, paradoxically, from the field of speaker identification. Sambur [47] has described a means of calculating from the regular LP parameters (or a non-linear transformation of them) a set of orthogonal parameters that divide into two groups: ones that vary significantly across the utterance, and ones essentially constant across it. He hypothesized that the first group reflects the linguistic features of the utterance, and the second its speaker dependent features. Use of the second group only for speaker identification gave excellent results.

The orthogonal parameters $\{\underline{b}(m)\}$ are calculated from any set of LP parameters $\{\underline{c}(m)\}$ as follows:

1. Calculate the covariance matrix $R = [r_{ij}]_{1(1)p}$ of the $\{\underline{c}(m) \mid m=1(1)M\}$ across the utterance:

$$r_{ij} = (1/M) \sum_{m=1}^{M} c_{im}c_{jm} - c_i c_j \qquad (3.41)$$

where $c_i = (1/M) \sum_{m=1}^{M} c_{im}$, and c_{im} = i th component of $\underline{c}(m)$.

2. Calculate the eigenvalues and eigenvectors of R by solving $|R-\lambda I|=0$ and $(R-\lambda_i I)\underline{e}_i=0$. Label the eigenvalues so that $\lambda_1>\lambda_2>\ldots>\lambda_p$, and scale the eigenvectors so that $\underline{e}_i^T\underline{e}_i = 1$.

3. Then the orthogonal parameters are

$$\underline{b}(m) = [\underline{e}_1 \ldots \underline{e}_p]^T \underline{c}(m)$$

or
$$\underline{b}(m) = E\,\underline{c}(m) \qquad (3.42)$$

and $\{b_{im}\}$, i=1(1)p' are the ones that vary significantly across the utterance. A suitable choice for p' is p/2. $E = [\underline{e}_1 \ldots \underline{e}_p]^T$ is an orthogonalizing matrix. Note that $E^T = E^{-1}$.

An orthogonal LP distance measure ORTHO

Since the $\{\lambda_i\}$ are the variances of the $\{b_i\}$ across the utterance, a natural measure of the dissimilarity between the test and reference is

$$\begin{aligned} d &= \sum_{i=1}^{p} (b_i - b_i')^2/\lambda_i \\ &= (E\underline{c}-E\underline{c}')^T \,\mathrm{diag}(1/\lambda_i)\, (E\underline{c}-E\underline{c}') \\ &= (\underline{c}-\underline{c}')^T E^T \mathrm{diag}(1/\lambda_i) E\, (\underline{c}-\underline{c}') \end{aligned}$$

by Eq. (3.42) if the $\{b_i\}$ and $\{b_i'\}$ are both derived from the same matrix E.

The CEPS distance measure of Eq. (3.30) can be written as $d_A(m,n) = \sqrt{2(\underline{c}-\underline{c}')^T(\underline{c}-\underline{c}')}$ where the $\underline{c}$'s are now the cepstral coefficients. This suggests that the ORTHO distance measure should be taken as

$$d_A(m,n) = \sqrt{k(\underline{c}-\underline{c}')^T W(\underline{c}-\underline{c}')} \qquad (3.43)$$

where k is a scaling constant, and the weighting matrix W is given by

$$W = E^T \text{diag}(1/\lambda_i)\, E \tag{3.44}$$

The orthogonalizing matrix E should be computed from a pooled covariance matrix R obtained from the individual cepstral covariance matrices of the test and reference utterances. The longer these utterances are the more stable will be the estimated eigenvalues and eigenvectors. Further, the implied summation in Eq. (3.43) should be carried out to the first p' terms only, to avoid including the speaker-identity-dependent elements of $\underline{b}$ and $\underline{b}'$.

CHAPTER 4

EXPERIMENTAL PROCEDURE AND RESULTS

4.1 DESCRIPTION OF EXPERIMENTAL WORK

A number of experiments were carried out to evaluate and improve the speech quality measures described in the previous chapter. A computer program was written to implement all of the proposed articulatory and prosodic quality measures with the exception of that for pitch. The measures were tested selectively on a variety of word pairs. The basic test was the comparison of two mono- or disyllabic words that differed in a single phoneme or prosodic feature.

It was decided to run the evaluation tests using pseudo quality defects, in which an "error" was the result of having a capable speaker deliberately mispronounce a word, rather than with real quality defects obtained from the speech of deaf or non-English speakers. This approach allows much more careful control over the errors, and makes it possible to investigate quality errors in the absence of interspeaker differences between test and reference utterances. A word list of 40 words was accordingly constructed that reflected the phonemic or prosodic errors most common among the deaf and ESL students. Where an English word was not available to provide a particular contrast, the appropriate nonsense word was used. The word list is shown in Fig. 4.1.

The list was read by four English Canadian speakers designated as JD, DK, RS (male), and EW (female). Each speaker read the list twice, with each reading taking about one minute. The speakers were instructed to articulate clearly, but otherwise to read normally. The readings were made in a quiet (acoustically screened) environment, and were recorded using a Bruel and

1	crystal	21	meat
2	thistle	22	mitt
3	this'll	23	mat
4	fuss	24	moot
5	fuzz	25	mot
6	bleating	26	might
7	bleeding	27	desert (v)
8	joy	28	desert (n)
9	zhoi	29	convict (v)
10	that	30	convict (n)
11	dat	31	object (v)
12	zat	32	object (n)
13	shin	33	I scream
14	chin	34	ice cream
15	win	35	hist'ry
16	wim	36	history
17	wing	37	eye
18	live	38	ah-ee
19	Liz	39	boy
20	riv	40	baw-ee

Fig. 4.1 Word list for speech quality tests

Kjaer condenser microphone type 4145 on a Scully 280 tape recorder operating at 38 cm/s.

The recorded speech was then bandpass filtered from 100 Hz to 4 kHz using a Krohn-Hite variable filter type 3342R, and was sampled at 10 kHz with a 12-bit analog-to-digital converter. A program was written to read this data and to semi-automatically segment it and eliminate the silent intervals between words. Endpoints were found by an algorithm examining the energy in each 10 ms frame of speech, but could be adjusted manually via a graphics display.

The main computer program, written in FORTRAN and run on the U.B.C. Computing Centre's Amdahl 470 V/6 Model II machine, then read the data corresponding to the desired word pair, and performed an LP analysis and speech quality comparison on it according to instructions given it. Output was in graphical form and was given via a Tektronix 4014 graphics terminal

or a hardcopy plot. The data acquisition and speech quality systems are represented in the flowcharts of Fig. 4.2.

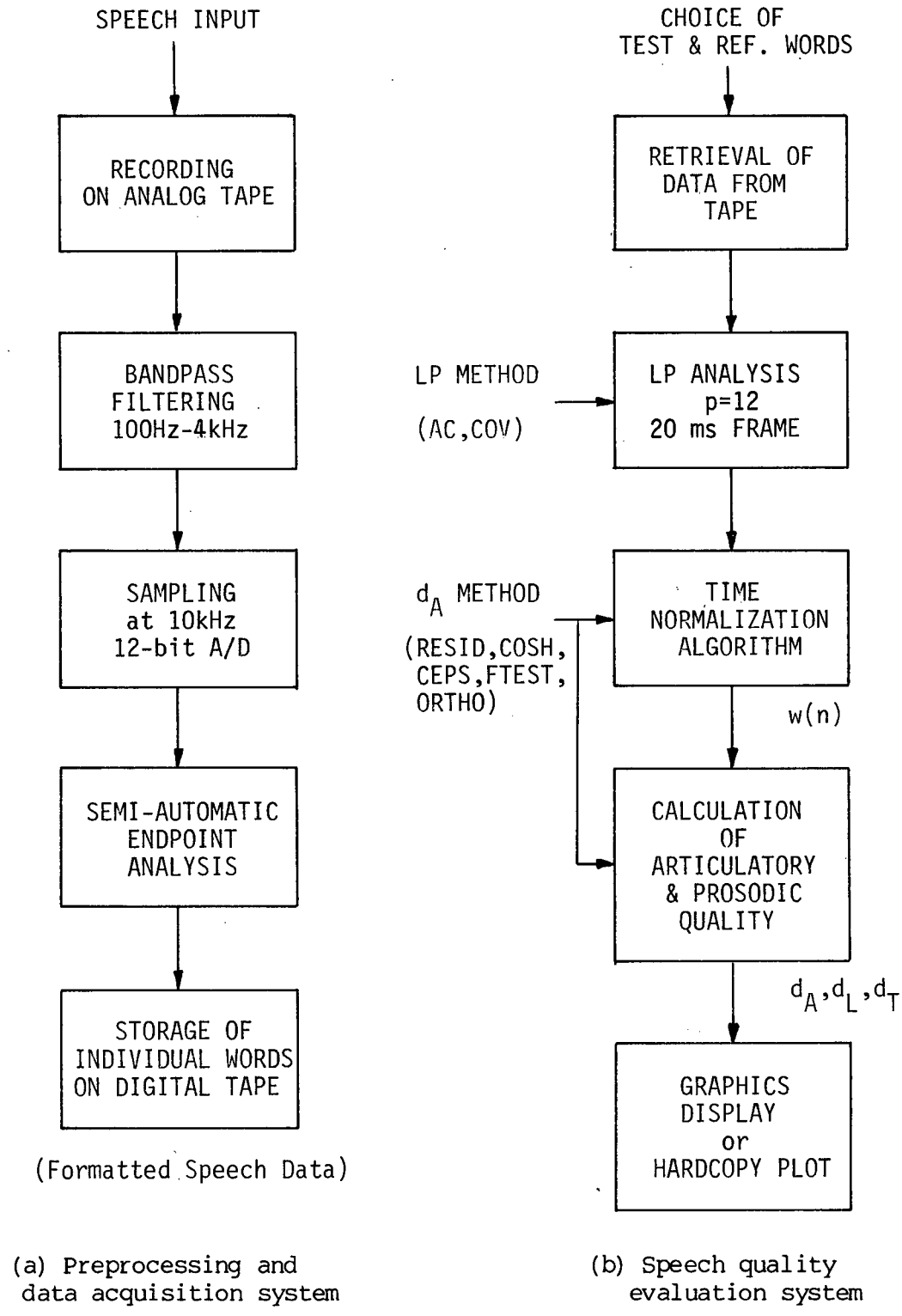


Fig. 4.2 Flowchart of speech processing system

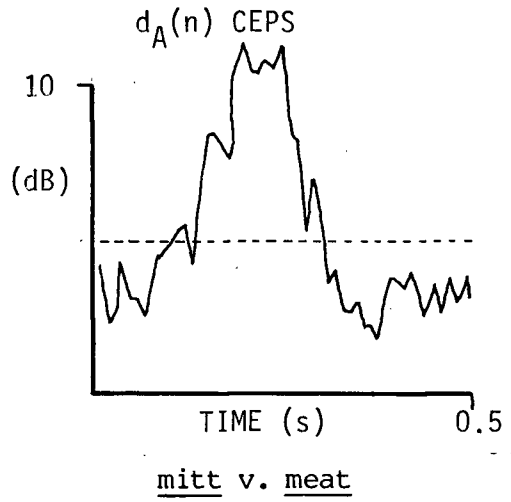
4.2 EVALUATION OF THE ARTICULATORY QUALITY MEASURES

In order to separate the question of sensitivity of the articulatory measures to quality errors from that of their susceptibility to interspeaker differences, tests were firstly run using test and reference utterances spoken by the same speaker. These tests are described in this section and the following one. It was felt that examining the effect of quality differences alone would help establish an upper bound on attainable performance.

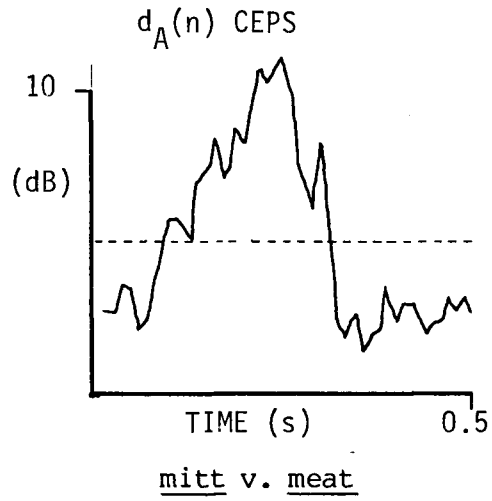
The first issue to settle was that of how the various articulatory quality measures performed relative to one another. Chapter 3 described two ways of calculating a set of linear prediction coefficients $\{a_k\}$ (AC and COV), and four ways of calculating a measure $d_A(m,n)$ for the quality difference between speech frames (RESID, COSH, CEPS, and FTEST). Initial tests sought to reduce this field of eight options. Only then was it feasible to investigate the greater issue, the capability or otherwise of one of the proposed schemes for detecting a variety of articulatory quality errors.

Comparison of AC and COV methods

It was found that the AC (autocorrelation) and COV (covariance) methods of linear prediction computed coefficients that differed significantly from one another - at times by up to 30 percent averaged across a speech frame - but which resulted in very similar articulatory quality measures. The differences in the quality measures were much less than either (i) the quality measure for identical words spoken by the same speaker but at different times, and (ii) the differences between quality measures for good and poor quality parts of the one word. This is illustrated in Fig. 4.3 which shows $d_A(n)$ for the comparison mitt v. meat, and LP data for frame 15



(a) AC (autocorr. method)



(b) COV (covariance method)

AC	a_k	COV
-0.09	a_1	-0.04
-0.64	a_2	-0.49
-0.48	a_3	-0.56
-0.42	a_4	-0.59
0.63	a_5	0.47
-0.04	a_6	0.10
0.04	a_7	0.21
0.32	a_8	0.30
-0.03	a_9	-0.05
0.31	a_{10}	0.26
-0.20	a_{11}	-0.19
-0.23	a_{12}	-0.21

(c) LP coefficients of meat after 0.15 s (at frame 15)

Fig. 4.3 Comparison of AC and COV methods of linear prediction

of meat computed by the AC and COV methods.

The unimportance as to whether the AC or COV method is used is rather surprising in view of the differences between coefficient sets. It implies that although the $\{a_k\}$ differ, they describe perceptually similar speech

spectra. For example, when meat (JD_1-AC)[1] is compared with meat (JD_1-COV), the average difference by the CEPS measure is only 1.18 dB, but when it is compared with meat (JD_2-AC), the difference rises to 3.48 dB.

The suitability of the AC method for cruder speech recognition experiments has been demonstrated repeatedly, but the results here show that use of the AC method costs little in sensitivity over the COV method even when fine comparisons of spectral similarity are required. Given the slightly better computational efficiency of the AC method, it follows that it is a better choice in practice. The remainder of this work, however, was carried out using the COV method. Interestingly, its supposed drawback of occasionally generating an unstable filter was never encountered.

Comparison of RESID, COSH, CEPS and FTEST measures

The various articulatory quality measures $d_A(m,n)$ were compared with one another on four pairs of test words. It was found that differences between the first three measures were small, and again relatively insignificant when compared to the variation in $d_A(m,n)$ across the words or the value of $d_A(m,n)$ for portions of good quality speech. The FTEST measure was an exception to this, giving results that were quite unrelated to those obtained using the other three. It sometimes responded very sharply to a quality error, but at other times failed to recognize one. Because it was computationally very expensive, it was not possible to run extensive tests with the method, but it does appear that FTEST is unsuitable as a reliable articulatory quality measure without some modification.

The RESID measure gave less consistent results than did the other two

---

[1] Symbols in parentheses specify which version of the preceding word is meant - in this case, the first recording of meat by JD with the AC method of linear prediction. Where no specification is given, the first recording by JD can always be assumed.

rms spectral measures, which performed very similarly. COSH weighted some quality differences more heavily than did CEPS, as expected from the discussion in [17]. The CEPS measure was preferred for subsequent tests, as it appeared to be the closest approximation to a true rms spectral measure, and because it offered superior computational efficiency.

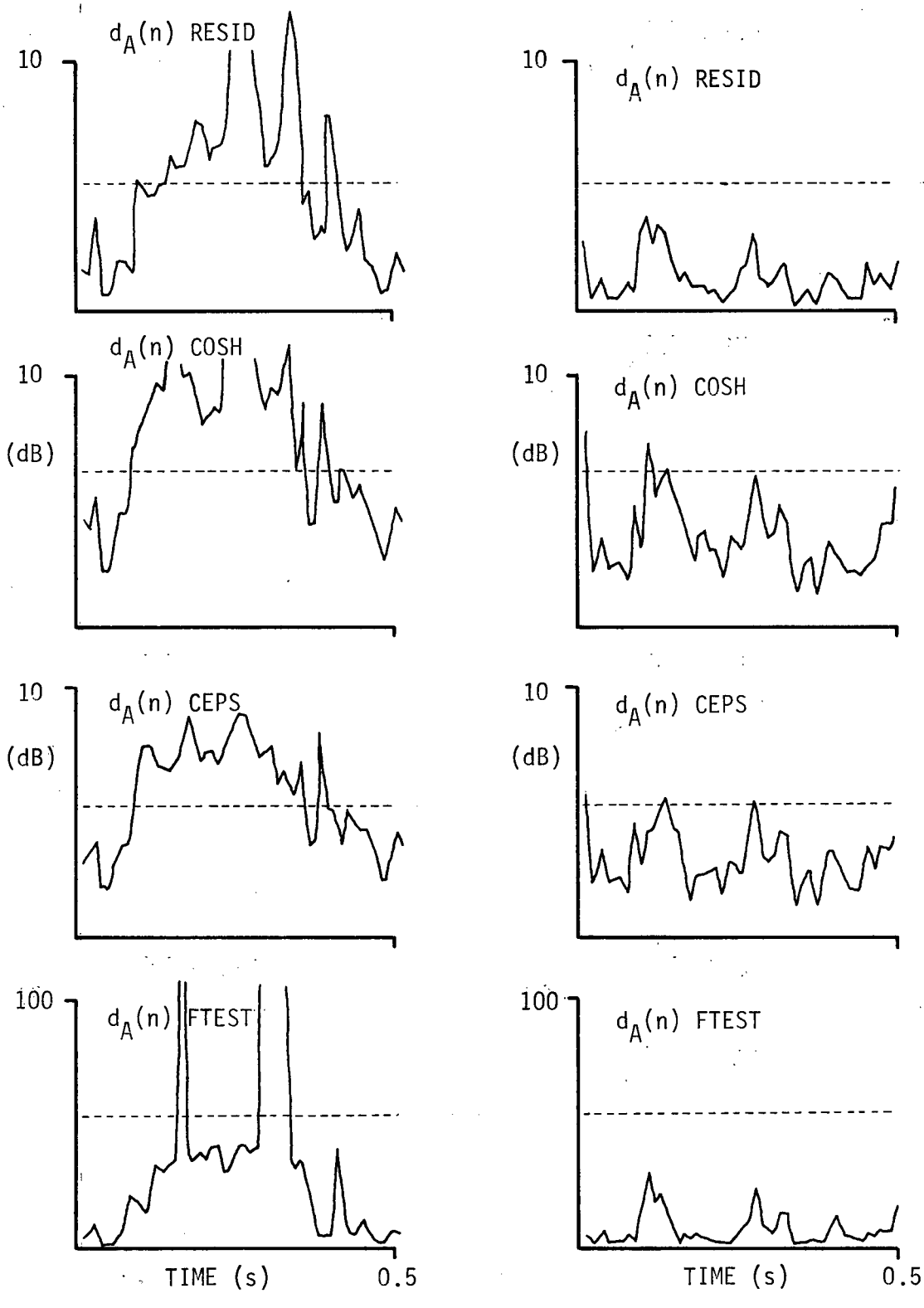
The four articulatory measures are compared with one another in Fig. 4.4 on the word pairs moot v. mat and live(2) v. live(1). All but FTEST are scaled by the factor 4.34.

The choice of articulatory measure has an effect on the time normalization function w(n) constructed by the dynamic programming algorithm, and hence on the prosodic timing measure. Despite differences in magnitude between the four articulatory measures, all resulted in very similar paths being taken by the DP algorithm. This was taken as evidence that the time normalization relationship between test and reference utterance was chosen correctly.

Ability to detect poor articulatory quality

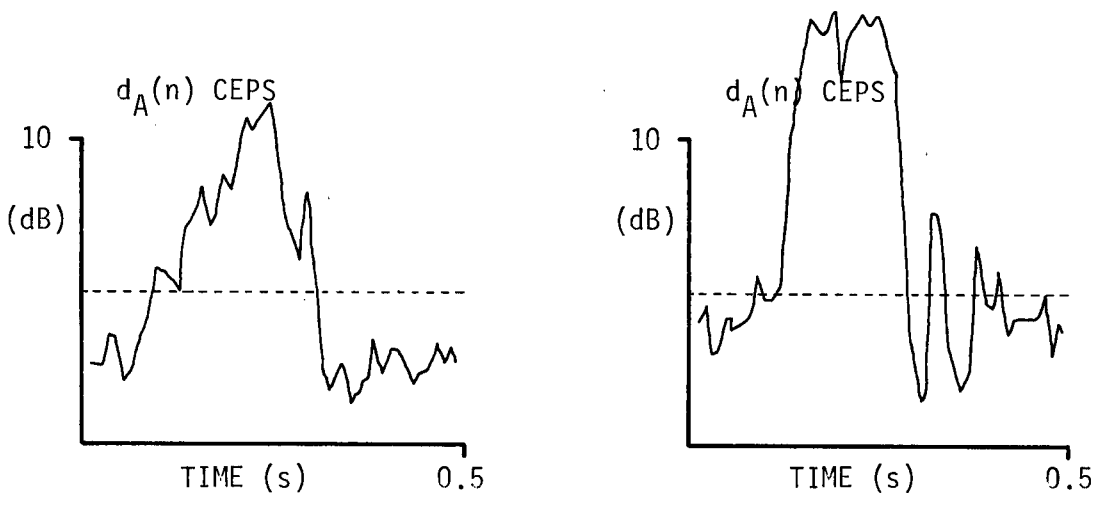
Contours of $d_A(n)$ were obtained for around 100 comparisons of test and reference words. The comparisons were done using the COV method to derive linear prediction data, and the CEPS measure for the actual comparison. A threshold of about 5 dB was found to generally indicate poor quality. The choice of test and reference words allowed a range of quality errors to be investigated, and the examples below give representative results for each category of articulation error. Appendix III gives details of the actual words compared in each category.

Most easily detected were errors in vowels (Fig. 4.5) and voiced fricatives and sononants (Fig. 4.6). Peak distances between word pairs were



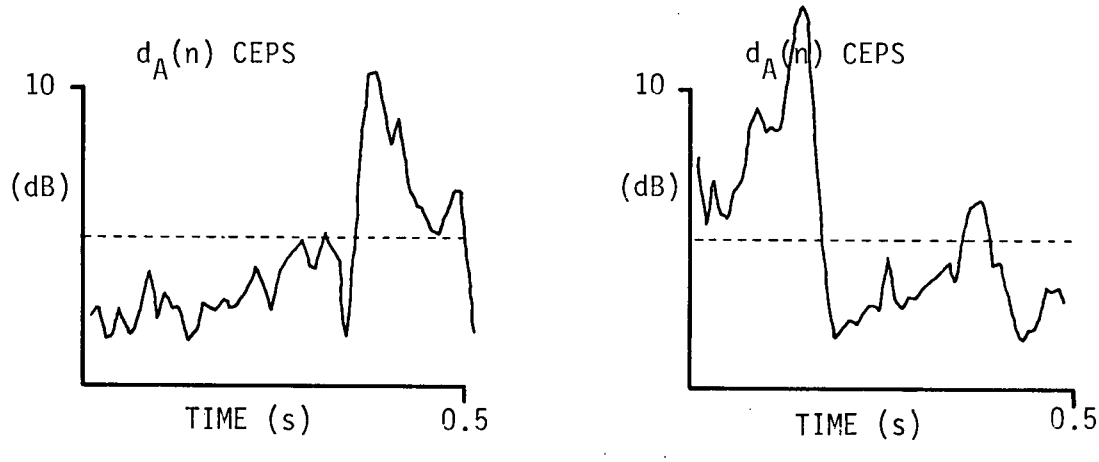
(a) moot v. mat (b) live(2) v. live(1)

Fig. 4.4 Comparison of the four articulatory measures



(a) mitt v. meat (b) mot v. meat

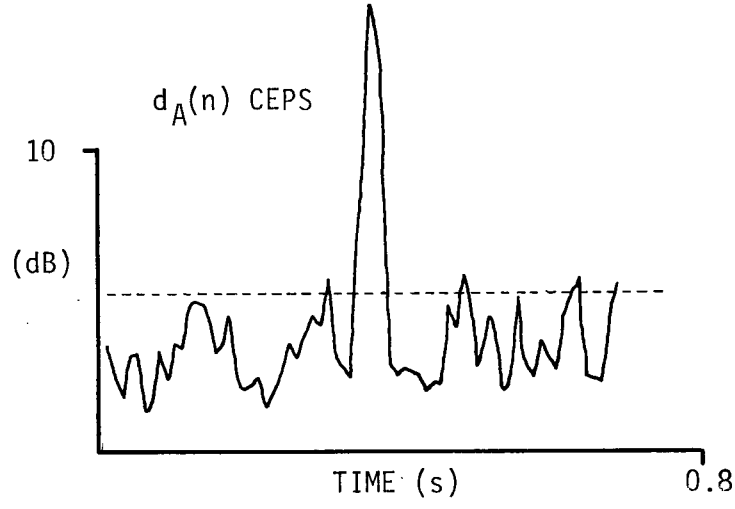
Fig. 4.5 Articulatory quality measure for vowel errors



(a) Liz v. live (b) riv v. live

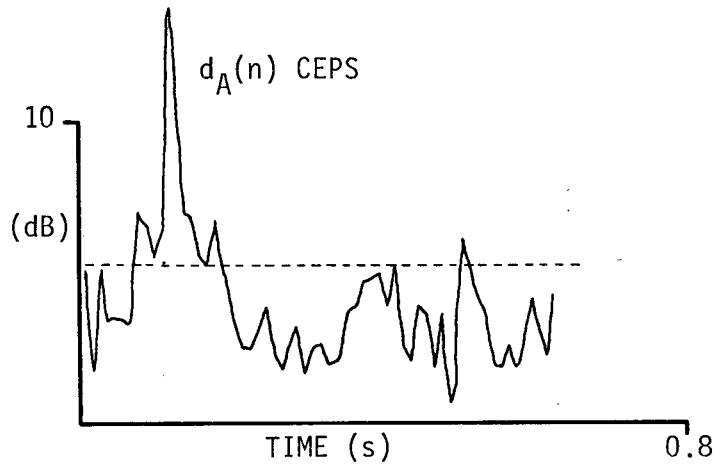
Fig. 4.6 Articulatory quality measure for voiced fricative and sonorant errors

in the range of 10 to 15 dB for the vowels, and 7 to 12 dB for the consonants. The strength of these sounds helped separate them from background noise, including quantization noise, and their prolonged repetitive structure appeared well suited for linear predictive analysis. Voicing errors had their greatest effect on the loudness measure, but where the error concerned a plosive, the characteristic puff of air called aspiration that is present after an unvoiced plosive was readily detected by the short duration peak in the articulatory measure (Fig. 4.7). Consonant substitutions involving plosives or fricatives could usually be detected (e.g. Fig. 4.8), but no characteristic patterns in the distance measure could be associated with them.



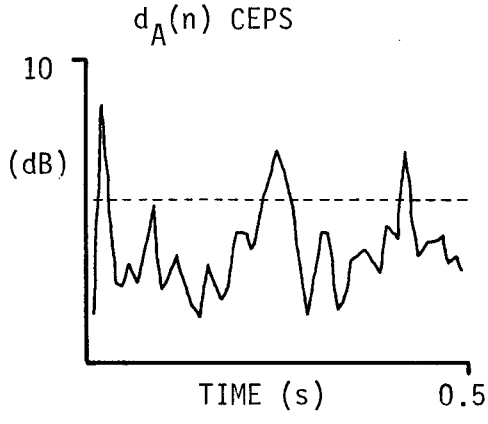
bleating v. bleeding

Fig. 4.7 Articulatory quality measure for plosive voicing errors

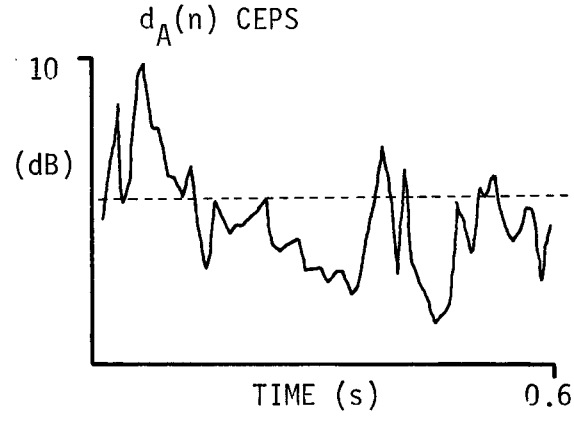


dat v. that

Fig. 4.8 Articulatory quality measure with other errors



(a) win v. wim



(b) mat v. that

Fig. 4.9 Articulatory quality measure for nasal errors

Errors in nasal sounds could not be detected at all in the word pairs investigated, though the substitution of a nasal for a fricative was apparent (Fig. 4.9). This points to a weakness of the all-pole linear prediction model of speech, which is not able to correctly model the zeros introduced by nasal coupling. The result is somewhat surprising though, for nasal sounds have been satisfactorily synthesized from an all-pole LP model [3], and there have been no indications in the literature that nasal sounds are particularly troublesome in speech recognition. But the test words used to check d_A for nasal sounds were very simple (wim-win-wing), and it is possible that with other words there would be greater coarticulation, which would assist both synthesis and recognition.

4.3 EVALUATION OF THE PROSODIC QUALITY MEASURES

Similiar tests to those of the previous section were made to investigate the prosodic quality measures of loudness and timing proposed in Chapter 3, and results of these are described below. Also given are experimental arguments for certain aspects of these measures, and for the modifications made to the dynamic programming time normalization procedure of Sakoe and Chiba. The development of the prosodic quality measures was much influenced by experimental results.

Time normalization path w(n)

The time normalization procedure of Sakoe and Chiba was found to choose a path through the network of (m,n) pairs that was, on a local level, erratic, and on a global level, occasionally quite wrong. That a chosen path correctly represents the actual time alignment between the two utterances is impossible to verify, but a grossly incorrect path can be identified by its unlikely shape. In a number of instances, the original algorithm chose a

path that indicated very rapid speech followed by very slow speech, when in fact the two words being compared were the same, spoken by the same speaker. Fig. 4.10 shows the phenomenon, with (a) following the incorrect path and (b) the path chosen by the modified algorithm.

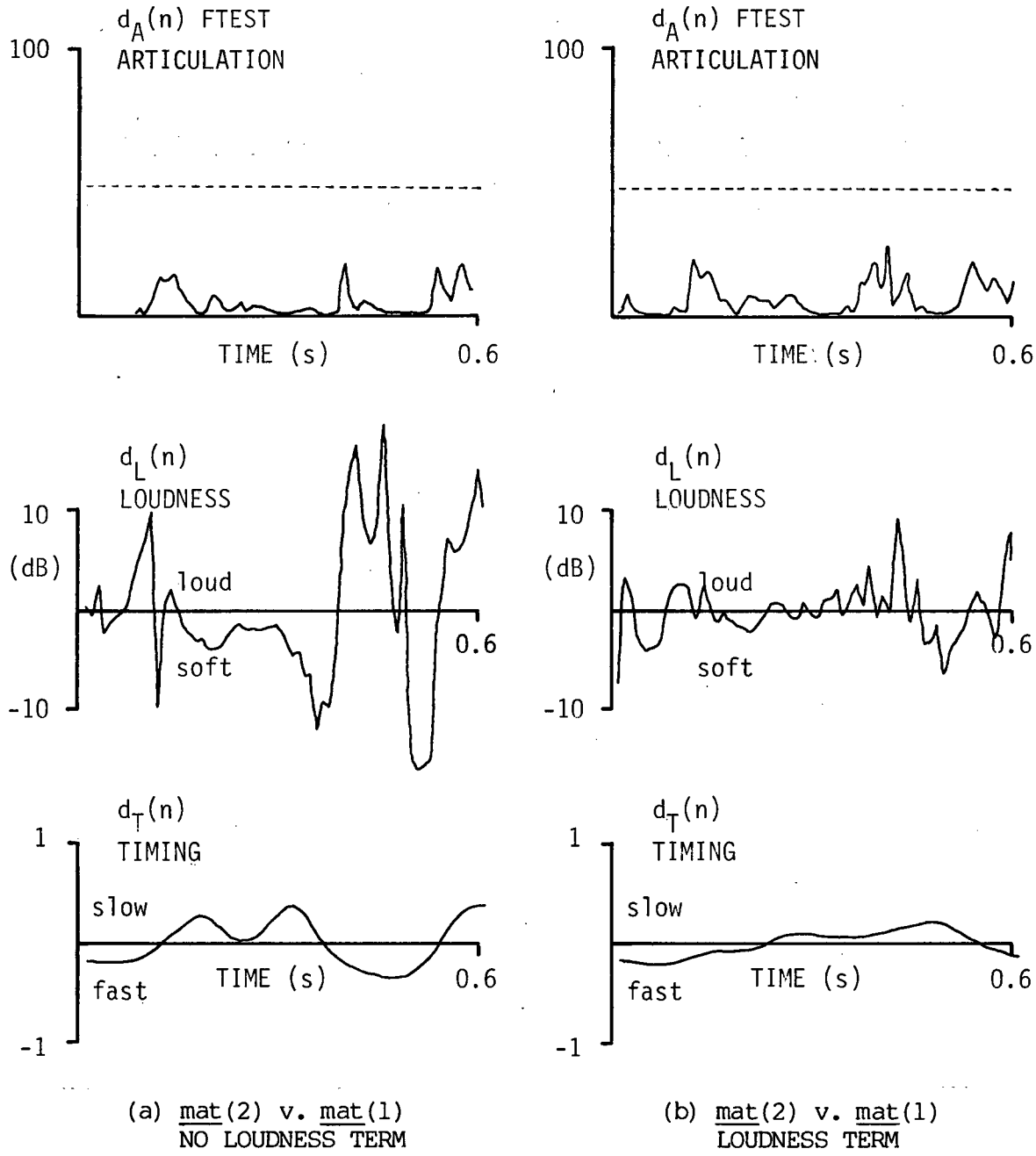


Fig. 4.10 Effect of loudness term in DP cost function

The algorithm was modified by the inclusion of an empirical loudness agreement term $0.2\,(L(m)-L'(n))^2$ in the cost function of the DP algorithm. The path chosen is therefore determined by agreement in loudness as well as in spectral shape. Even better than the term $(L-L')^2$ would be a function of the loudness measure itself, e.g. $d_L^2 = (a_1L+a_2-L')^2$, but as a_1 and a_2 can be calculated only after the path is found, this would require iterative computation which cannot be justified in terms of computing effort.

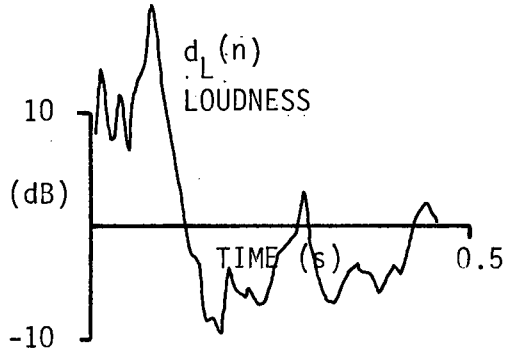
Local irregularities in w(n) were felt to be without physical meaning, and a method was sought of eliminating them. It was accomplished by restricting the speed with which the algorithm could switch the rate of increase of w(n) between its maximum and minimum allowable values (2 and 0 respectively). This was done by imposing the conditions that

$$\begin{aligned} \Delta w(n) &\neq 2 \quad \text{if} \quad \Delta w(n-1) = 0 \\ \Delta w(n) &\neq 0 \quad \text{if} \quad \Delta w(n-1) = 2 \end{aligned} \tag{4.1}$$

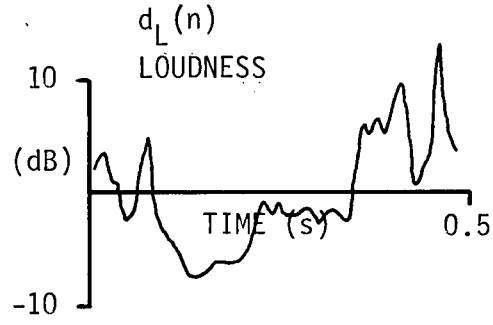
where $\Delta w(n)$ denotes $w(n)-w(n-1)$ etc. These additional restrictions had a negligible effect on the assessed articulatory quality, but resulted in a smoother w(n).

The loudness quality measure d_L

The loudness measure was found to respond strongly to both voicing errors and syllable stress errors, as seen from Fig. 4.11. For voicing errors, the surprising result was obtained that unvoiced sounds frequently have greater loudness than the corresponding voiced ones. However, the loudness measure was greatly affected by other aspects of the speech, especially vowel errors, and identification of the above errors is difficult without prior knowledge about them. Indeed, the loudness quality measure d_L must be judged a rather unreliable indicator of true loudness quality.

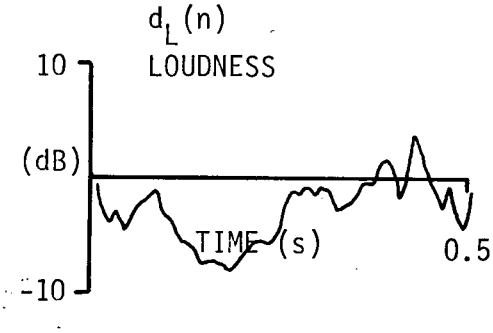


(a) thistle v. this'll (DK_1)
(Voicing error)

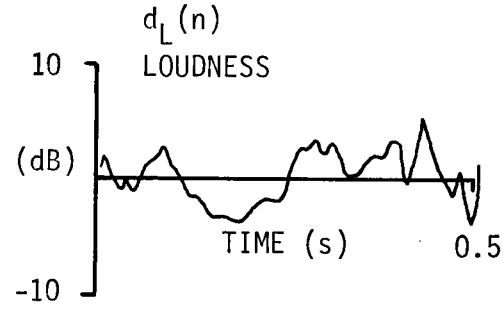


(b) con-vict'(v) v. con'-vict(n)
(Syllable stress error)

Fig. 4.11 Loudness quality measure for voicing and syllable stress errors



(a) Liz v. live
NO CORRECTION



(b) Liz v. live
CORRECTION
a_1=1.13, a_2=5.6 dB

Fig. 4.12 Loudness quality measure: effect of correction factors

Its main use is likely to be in providing visual feedback during speech training of the magnitude of a particular error; in its present form it is not really suitable as a diagnostic tool.

The effect of the correction factors a_1 and a_2 was examined. These factors attempt to compensate for differences between the test and reference utterances of average loudness and range of variation of loudness. Fig. 4.12 shows a contour for d_L with and without the correction factors included. The factors are likely to prove most useful in recording situations less carefully controlled than the one here.

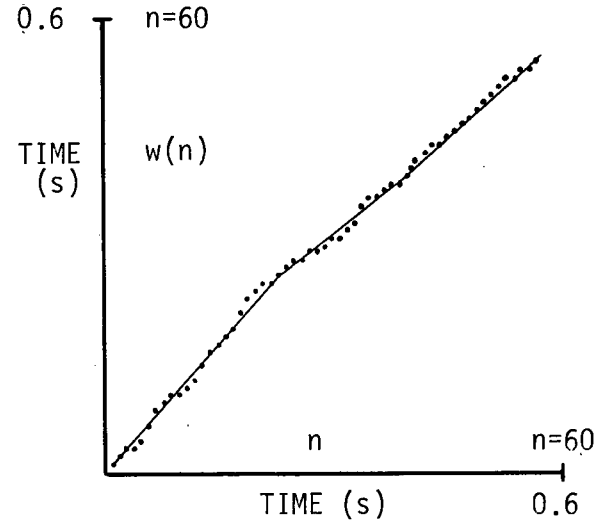
The timing quality measure d_T

Derivation of a suitable timing measure required finding a satisfactory definition of w'(n), the rate of change of the time alignment function w(n). Because of the integerized nature of w(n), the usual algorithms of numerical analysis do not yield a sufficiently smooth w'(n). After some experimentation it was found that good results could be obtained using the cubic spline curve smoothing algorithm of Reinsch [44],[45], and taking w'(n) to be the slope at n of the smoothed function. Reinsch's algorithm finds the function having minimum average squared second derivative (hence: a cubic spline) among all functions $w^*(n)$ satisfying

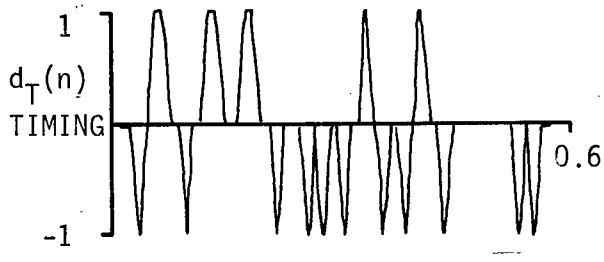
$$(1/N) \sum_{n=1}^{N} (w(n)-w^*(n))^2 = S \qquad (4.2)$$

where S is a constant controlling the degree of smoothing; S=1/2 was found to give the most satisfactory smoothing.

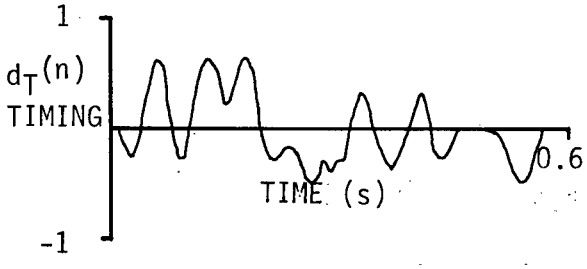
Fig. 4.13 shows the resultant timing quality measure $d_T(n)$ for three methods of calculating w'(n). It is seen that the Lagrangian difference formulas perform very poorly indeed.



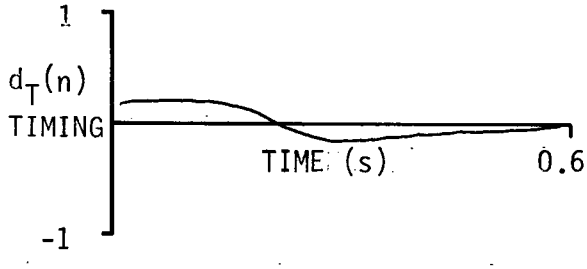
(a) Time alignment function w(n) for dat v. that



(b) Using 2-point Lagrangian differentiation



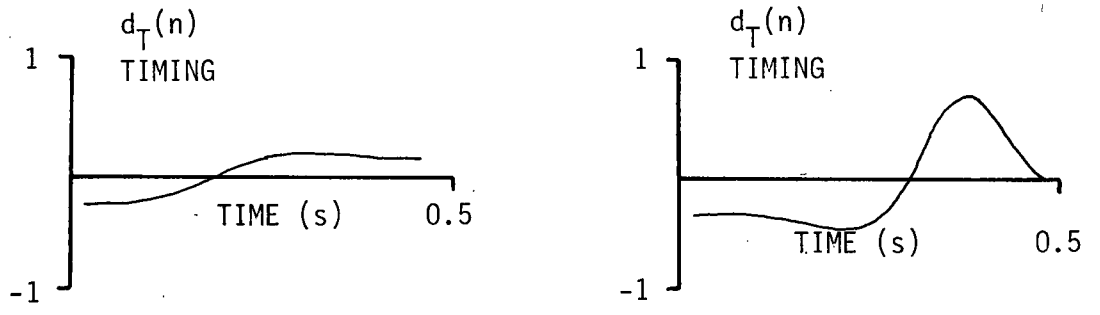
(c) Using 5-point Lagrangian differentiation



(d) Using cubic spline smoothing (S=1/2)

Fig. 4.13 Timing quality measures for differently computed w'(n)

The timing measure was found to give a useful indication of speed variation within a word, as can be seen by the results in Fig. 4.14 in which pairs of words having voicing and syllable stress contrasts are compared. However, variations in speed comparable to those obtained in such situations were found to occur in other instances too. These variations were partly due to articulatory errors, and partly due to entirely natural variations in speech. They overshadowed, for instance, the variations due to omission of a syllable or due to a diphthongization error. Though it appears to faithfully track speed variations in speech, the timing quality measure too must be regarded as useful primarily for producing visual feedback during corrective speech exercises, rather than for diagnostic purposes.



(a) thistle v. this'll (DK_1) (Voicing error)

(b) con-vict'(v) v. con'-vict(n) (Syllable stress error)

Fig. 4.14 Timing quality measure for voicing and syllable stress errors

4.4 EFFECT OF INTERSPEAKER DIFFERENCES

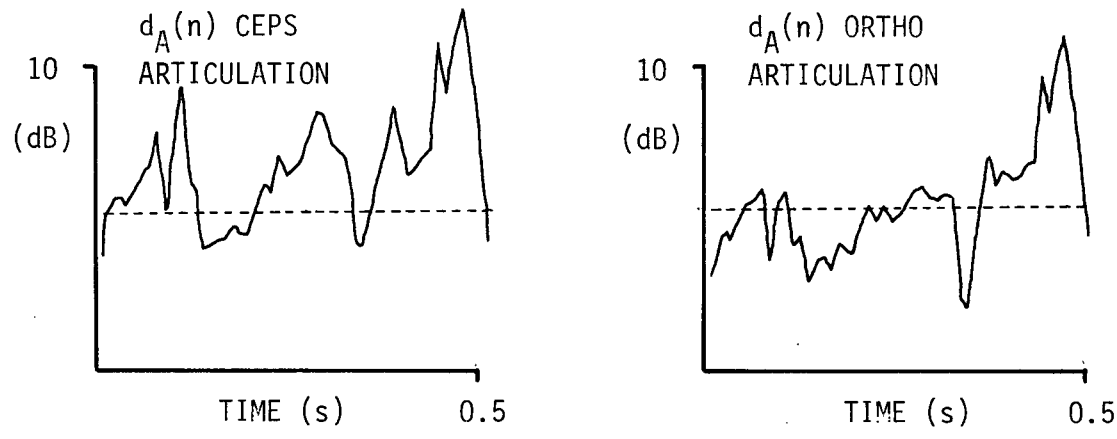
Tests were made with several combinations of speakers and word pairs to determine the deleterious effect, if any, of interspeaker differences on the various speech quality distance measures. The performance of the ORTHO (orthogonal linear prediction) distance measure of Section 3.3 was also examined.

Articulatory quality

Deterioration in the deduced speech quality was definitely noticed with the CEPS measure. The level of 'background' disagreement (i.e. $d_A(n)$ for sections of good quality) increased by about 3 - 4 dB, and peaks in the distance measure of the order of 10 dB were found to occur in passages where there were no differences in the articulated speech sounds. These phenomena may be observed in the example of Liz(DK_1) v. live(JD_1), shown in Fig. 4.15(a). The improvement sometimes achievable with the ORTHO distance measure is shown in Fig. 4.15(b) (scaling factor k = 0.05 x 4.34 dB), where the false peaks in d_A alone are reduced. However, this improvement was not always obtained, and in about 30 per cent of the cases examined, even the ORTHO articulatory measure implied a quality error where there was none. (Actual errors were always indicated, to about the same degree as for no interspeaker differences.) The ORTHO measure, therefore, has some use in reducing interspeaker articulatory differences, but it is not a final solution. It is possible that an increase in utterance length for calculation of the covariance matrix would bring further improvements, but these are likely to be minor.

Prosodic quality

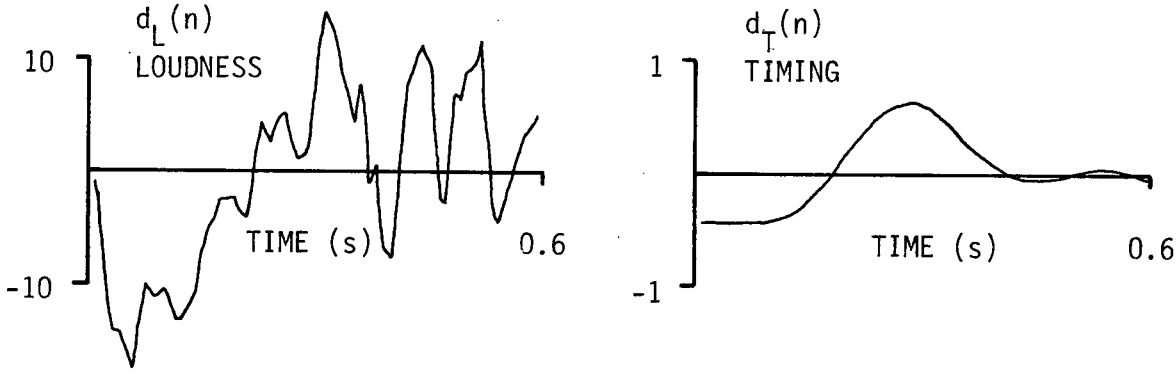
Despite the errors in the articulatory quality function, the dynamic programming algorithm continued to choose a reasonable time alignment path



(a) Liz(DK_1) v. live(JD_1)
CEPS measure

(b) Liz(DK_1) v. live(JD_1)
ORTHO measure

Fig. 4.15 Articulatory quality measures for interspeaker differences



(a) ob-ject'(RS_1) v. ob'-ject(JD_1)
Loudness measure $d_L(n)$

(b) ob-ject'(RS_1) v. ob'-ject(JD_1)
Timing measure $d_T(n)$

Fig. 4.16 Prosodic quality measures for interspeaker differences

between the test and reference utterances, indicating that relationships between adjacent (m,n) sample points remain well preserved in the presence of interspeaker differences. Whether the CEPS or ORTHO distance measure was used made very little difference to the path chosen by DP algorithm. Fig. 4.16 shows the loudness and timing distance measures obtained for the case of reversed syllable stress for different speakers. It is seen that both measures remain meaningful indicators of prosodic quality within the constraints discussed previously.

Interspeaker differences, then, have an adverse effect on the articulatory quality measure, but not on the prosodic quality measures. Useful evaluations of articulatory quality are still possible, but false errors are often indicated. It is important to investigate further ways of reducing interspeaker differences, and these may have to involve the inclusion of information not obtainable from the linear prediction parameters alone.

CHAPTER 5

CONCLUSIONS

This thesis is concerned with an investigation of the feasibility of automatic speech quality analysis. A computer-based system that can assess the speech quality of an input utterance will have application in speech training of the deaf and of second language students, and will partly integrate the special-purpose devices existing now.

A speech pathologist's view was taken of speech quality, which was regarded as the lack of defects in the components of speech - voice, articulation, and prosody. A set of quality measures, based on the all-pole linear prediction model of speech, was proposed for expressing the articulatory and prosodic quality between a pair of utterances.

Evaluations made of the measures and of aspects of linear prediction showed firstly that the difference between the autocorrelation and covariance methods of linear prediction was not significant for speech quality analysis. The differences between results with the RESID, COSH, and CEPS measures were also slight, but the CEPS (cepstral) measure was preferred because of its theoretical accuracy and computational efficiency. The proposed FTEST measure gave inconsistent results, and was rejected in its current form.

The CEPS measure was found to be effective in detecting most of the common errors of articulation, with the exception of errors between nasal sounds. A general threshold for deciding between good and poor quality was 5 dB. Vowel errors registered peak disagreements of up to 15 dB, and voiced fricative and sonorant errors peaks up to 12 dB.

A valuable indicator of prosodic quality was derived from the time alignment function w(n) - a by-product of the dynamic programming algorithm

for matching the test and reference utterance time axes with one another. The timing measure was most effective in showing errors in syllable stress and voicing function, and appeared to be accurate in tracking speed variations in general. The loudness measure also responded clearly to these errors. Neither of these measures, however, was particularly satisfactory in diagnosing such an error (or other errors of prosody), and both will likely find most use in the monitoring of error magnitudes.

Interspeaker differences did occasionally mask articulatory errors, and indicate poor quality where there was none. The ORTHO measure, derived from orthogonalized cepstral coefficients, cancelled these differences to a degree, but not always sufficiently. The prosodic quality measures were relatively immune to interspeaker differences, in part because such speaker-dependent properties as average value and dynamic range of a quantity are removed in the definition of the measures.

Work remains to be done in several key areas. Distance measure data needs to be collected over a full range of quality errors, to allow suitable functions f_i to be found for the overall quality measures. These functions are dependent on knowledge about the amounts of variation in the $d_i(n)$ that are normal or else indicative of an error. Decisions need to be made as to appropriate display modalities for the computed quality measures, and these will be influenced by the actual teaching program designed for use with the system as a training aid.

The work begun on interspeaker differences will have to be extended. Larger interspeaker differences will be encountered in practice than were examined here, and poorer performance of the quality measures can be expected. Although additional improvement may be obtained by continuing in the directions of this work, it is likely that new methods will have to be

developed. One idea is to cancel the effect of differences in vocal tract lengths by transformations of the filter coefficients. Another is to make use of articulatory models such as Coker's [9] to characterize and then compensate for the differences between learned motions of the articulators. The problem of interspeaker differences is actually much more tractable for speech quality analysis than it is for speech recognition. It is quite acceptable to require student and teacher to initialize the system by speaking a standard sentence from which their individual characteristics can be identified, and even to require input of the phonemic representation of the speech being evaluated.

Other areas for further research include implementation and testing of the proposed pitch measure, together with investigation of the required accuracy for the pitch detector; adoption of a more general linear prediction model that will allow nasal zeros to be represented exactly; and removal of the constraints of fixed endpoints and maximum slope range of 1/2 to 2 in the time normalization algorithm. Rabiner et al. have reported some algorithms for this in [41]. Allowing a greater slope range will be important if the speech quality system is to be used with deaf children, who tend to speak 2 to 4 times more slowly than normal speakers.

In spite of these needed extensions, the results of this preliminary investigation suggest that automatic speech quality analysis by computer is practical. Such computer analysis of speech may one day find useful application in speech training.

APPENDIX I

THE PHONETIC ALPHABET FOR ENGLISH

These tables give the symbols of the International Phonetic Alphabet for the sounds occurring in English. The classification used is described in Section 2.1 of the main text. More detailed classifications and variations in symbology are also possible, see for example [54], [31].

A. Vowels and diphthongs

DEGREE OF CONSTRICTION	TONGUE HUMP POSITION front		center		back		DIPHTHONGS	
high	/i/	ē	/ɝ/	ûr	/u/	o͞o		
	/I/	i	/ə˞/	ər	/U/	o͝o	/aI/	ī
							/ɔI/	oi
medium	/e/	ā	/ɜ/	û	/o/	ō	/aU/	ow
	/ɛ/	e	/ə/	ə	/ɔ/	aw	/ju/	ū
			/ʌ/	u				
low	/a/	a			/ɒ/	o		
	/æ/	a			/ɑ/	ah		

B. Consonants

PLACE OF ARTICULATION	MANNER OF PRODUCTION fricatives		plosives		sonorants		nasals	
bilabial			/p/	p	/ʍ/	wh	/m/	m
			/b/	b	/w/	w		
labiodental	/f/	f						
	/v/	v						
dental	/θ/	th						
	/ð/	dh						
alveolar	/s/	s	/t/	t	/l/	l	/n/	n
	/z/	z	/d/	d				
palatal	/ʃ/	sh	/tʃ/	ch	/j/	y	/ŋ/	ng
	/ʒ/	zh	/dʒ/	j	/r/	r		
velar			/k/	k				
			/g/	g				
glottal	/h/	h						

Notes

1. Each phonetic symbol is followed by its usual dictionary transcription. Note the pronunciation of û (fur), o͞o (boot), o͝o (foot), th (thin), dh (this), zh (azure).

2. Vowels: /ə/, /ə̂/ are unstressed. /ɝ/, /ɑ/ are the equivalent in the General American accent of /ɜ/, /ɒ/. /a/ is used in New England and elsewhere in place of /æ/. /u/, /U/ have no initial form, and /I/, /ɛ/, /a/, /æ/, /ʌ/, /U/, /ɒ/, /ɑ/ have no final form.

3. Consonants: The groupings represent unvoiced-voiced pairs, with the exception of /j/-/r/ which are unrelated. /tʃ/, /dʒ/ are actually affricates, not plosives. /ʒ/, /ŋ/ have no initial form, and /h/, /ʍ/, /w/, /j/ have no final form.

APPENDIX II

ALGORITHMS

This Appendix gives details for several of the algorithms mentioned in Chapter 3.

1. Solving the LP autocorrelation equations

The Levinson method is an elegant recursive solution to the p linear equations in $\{a_k\}$ given in Eq. (3.6):

$$\sum_{k=1}^{p} a_k r_{|i-k|} = - r_i, \qquad i=1(1)p$$

the method is derived in [3], [11], and [29]. As stated below it is due to Makhoul [25].

1. Put $\alpha_0=r_0$
2. For i=1(1)p evaluate

$$a_i^{(i)} = (-1/\alpha_{i-1})(r_i + \sum_{k=1}^{i-1} a_k^{(i-1)} r_{i-k})$$

$$a_k^{(i)} = a_k^{(i-1)} + a_i^{(i)} a_{i-k}^{(i-1)} \qquad \text{for } k=1(1)i-1$$

$$\alpha_i = (1-a_i^{(i)2})\alpha_{i-1}$$

3. Then $a_i=a_i^{(p)}$, i=1(1)p, and $\alpha=\alpha_p$

The $a_i^{(i)}$ are identical with the reflection (parcor) coefficients k_i.

2. Calculating the r_k from the a_k

The Levinson method gives a procedure for transforming the autocorrelation coefficients $\{r_k\}$ into the filter coefficients $\{a_k\}$. Sometimes it is necessary to make the reverse transformation, i.e. to find

the $\{r_k\}$ that satisfy Eq. (3.6) for a given set of $\{a_k\}$.

The algorithm is derived in [3], and involves firstly computing the $\{a_k^{(i)}\}$ and then finding the $\{r_k\}$ from these. The computed $\{r_k\}$ are normalized with respect to r_0, i.e. $r_0=1$.

1. $a_k^{(p)} = a_k$ for k=1(1)p

   For i=p(-1)2 do: for k=1(1)i-1 do: evaluate

$$a_k^{(i-1)} = (a_k^{(i)} - a_i^{(i)} a_{i-k}^{(i)})/(1-a_i^{(i)2})$$

2. $r_1 = -a_1^{(1)}$

   For i=2(1)p evaluate

$$r_i = - a_i^{(i)} - \sum_{k=1}^{i-1} r_k a_{i-k}^{(i)}$$

3. Efficient calculation of the covariance matrix

The COV method requires calculation of $(p+1)^2/2$ covariances, defined by Eq. (3.11):

$$\phi_{ik} = \phi_{ki} = \sum_{n=p}^{N-1} x(n-i)x(n-k), \qquad i,k=0(1)p$$

An efficient way of calculating these covariances, for $N \gg p$, is to calculate only ϕ_{i0}, i=0(1)p, from the definition, and then to make use of the relationship

$$\phi_{ik} = \phi_{i-1,k-1} + x(p+1-i)x(p+1-k) - x(N+1-i)x(N+1-k)$$

for i=1(1)p, k=1(1)i. This enables calculation of the covariances to be almost as fast as calculation of the autocorrelations defined in Eq. (3.7).

4. Dynamic programming to find the optimal w(n) (after Itakura [22])

The optimal time warping funtion w(n) is defined here to be the mapping that minimizes the total distance $D = \sum_{n=1}^{N} d_A(w(n),n)$, subject to the endpoint and continuity constraints

1. $w(1)=1$, $w(N)=M$
2. $w(n+1)-w(n) = 0$, 1, or 2 if $w(n)\neq w(n-1)$, else 1 or 2.

Let D(m,n) represent the optimum distance from (1,1) to (m,n), so that D(M,N) = D. Dynamic programming then makes use of the relationship

$$D(m,n) = \min_{m'} (D(m',n-1)) + d(m,n)$$

The complete algorithm, incorporating the constraints on w(n), is as follows:

1. Put $D(1,1)=d(1,1)$, $h(1,1)=1$.

   Define for n=1(1)N:

   $$m_L(n) = \max([(n+1)/2], 2n+M-2N)$$

   $$m_U(n) = \min(2n-1, [(n+1+2M-N)/2])$$

2. For n=2(1)N do: for $m=m_L(n)(1)m_U(n)$ do:

   (a) Find the m' in the range $\max(m_L(n-1), m-2)$ to $\min(m_U(n-1), m)$ - but excluding m'=m if h(m,n-1)=0 - for which D(m',n-1) is a minimum.

   (b) Put $D(m,n) = D(m',n-1) + d(m,n)$

   $h(m,n) = m-m'$

3. Then D = D(M,N), and w(n) is found via the recursion:

   $w(N) = M$

   $w(n-1) = w(n) - h(w(n),n)$, n=N(-1)2.

APPENDIX III

LIST OF WORD PAIRS COMPARED

The following list sets out the full selection of word pair comparisons made in obtaining the results of Chapter 4.

1. Comparing AC and COV methods:

mitt/meat (AC/COV); riv/live (AC/COV); [kit/cat (RE: AC/COV)].

2. Comparing RESID, COSH, CEPS, FTEST measures:

moot/mat (all 4); live(2)/live(1) (all four); zat/dat (all 4); Liz/live (all 4).

3. Evaluating articulatory quality:

mitt/meat ($JD_1/JD_2/DK_1/RS_1/EW_1$); mat/meat; moot/meat; mot/meat; might/meat; moot/mat; Liz/live (JD_1/RE); riv/live (JD_1/RE); bleating/bleeding ($JD_1/DK_1/EW_1$); dat/that; zat/that; zat/dat; joy/zhoi; shin/chin; win/wim; mat/that; wing/wim; mat/zat.

4. Evaluating prosodic quality:

dat(2)/dat(1) (FTEST); [cat(2)/cat(1) (RE: FTEST)]; Liz/live; mitt/meat (DK_1); thistle/this'll ($JD_1/JD_2/DK_1/EW_1$); fuss/fuzz ($JD_1/JD_2/DK_1$); convict/convict; desert/desert ($JD_1/JD_2/DK_1/EW_1$); object/object ($JD_1/DK_1/RS_1/EW_1$).

5. Interspeaker differences:

Liz/live ($JD_1/DK_1/DK_1$-JD_1: CEPS/ORTHO); mitt/meat ($JD_1/DK_1/RS_1/EW_1/DK_1$-JD_1/JD_1-DK_1/JD_1-RS_1/EW_1-JD_1: CEPS/ORTHO); object/object ($JD_1/RS_1/RS_1$-JD_1: CEPS/ORTHO); bleating/bleeding (DK_1-JD_1).

BIBLIOGRAPHY

[1] B.S. ATAL, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". J. Acoust. Soc. Amer. 55: 1304-1312, June 1974.

[2] B.S. ATAL, "Automatic recognition of speakers from their voices". Proc. IEEE 64: 460-475, April 1976.

[3] B.S. ATAL & S.L. HANAUER, "Speech analysis & synthesis by linear prediction of the speech wave". J. Acoust. Soc. Amer. 50(2): 637-655, 1971.

[4] A. BOOTHROYD, "Some experiments on the control of voice in the profoundly deaf using a pitch extractor and storage oscilloscope display". IEEE Trans. Audio & Electroac. AU-21: 274-278, June 1973.

[5] I.P. BRACKETT, "Parameters of voice quality". In Travis [55] (1971), 441-464.

[6] D.R. CALVERT, "Some acoustic characteristics of the speech of profoundly deaf individuals". Ph.D. dissertation, Stanford Univ., Calif., 1961.

[7] D.R. CALVERT, "Deaf voice quality: A preliminary investigation". Volta Rev. 64: 402-403, 1962.

[8] R. CATHART (ed.), Human communication and its disorders. U.S. Dept. of Health, Education, and Welfare, 1969.

[9] C.H. COKER, "A model of articulatory dynamics and control". Proc. IEEE 64: 452-460, April 1976.

[10] L.E. CONNOR (ed.), Speech for the deaf child: knowledge and use. A.G. Bell Ass. for the Deaf, Washington D.C., 1971.

[11] R.G. CRICHTON & F. FALLSIDE, "Linear prediction model of speech production with applications to deaf speech training". Proc. IEE 121: 865-873, Aug 1974.

[12] H. DAVIS & S.R. SILVERMAN (eds.), Hearing and deafness, 4th ed. Holt, Reinhart, & Winston, New York, 1978.

[13] P.V. de SOUZA, "Statistical tests & distance measures for LPC coefficients". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-25: 554-559, Dec 1977.

[14] G.C.M. FANT, Acoustic theory of speech production. Moulton, The Netherlands, 1960.

[15] J.L. FLANAGAN, Speech analysis, synthesis, and perception, 2nd ed. Springer-Verlag, Berlin, 1972.

[16] D.J. GOODMAN et al., "Intelligibility and ratings of digitally coded speech". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-26: 403-409, Oct 1978.

[17] A.H. GRAY, Jr., & J.D. MARKEL, "Distance measures for speech processing". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-24: 380-391, Oct 1976.

[18] V.N. GUPTA, J.K. BRYAN, & J.N. GOWDY, "A speaker-independent speech-recognition system based on linear prediction". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-26: 27-33, Feb 1978.

[19] C.V. HUDGINS, "A comparative study of the speech coordination of deaf and normal subjects". J. Genet. Psychol. 44: 1-48, 1934.

[20] C.V. HUDGINS & F.C. NUMBERS, "An investigation of the intelligibility of the speech of the deaf." Genet. Psychol. Monograph 25: 289-392, 1942.

[21] IEEE Recommended Practice for Speech Quality Measurements (Standards publication no. 297), IEEE Trans. Audio & Electroac. AU-17: 227-246, Sep 1969.

[22] F. ITAKURA, "Minimum prediction residual principle applied to speech recognition". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-23: 67-72, Feb 1975.

[23] D.N. KALIKOW & J.A. SWETS, "Experiments with computer-controlled displays in second-language learning". IEEE Trans. Audio & Electroac. AU-20: 23-28, March 1972.

[24] H. LEVITT, "Speech processing aids for the deaf: an overview". IEEE Trans. Audio & Electroac. AU-21: 269-273, June 1973.

[25] J. MAKHOUL, "Linear prediction: a tutorial review". Proc. IEEE 63: 561-580, April 1975.

[26] J.D. MARKEL, "Digital inverse filtering - a new tool for formant trajectory estimation". IEEE Trans. Audio & Electroac. AU-20: 129-137, June 1972.

[27] J.D. MARKEL, "The sift algorithm for fundamental frequency estimation". IEEE Trans. Audio & Electroac. AU-20: 367-377, Dec 1972.

[28] J.D. MARKEL & A.H. GRAY, Jr., "On autocorrelation equations as applied to speech analysis". IEEE Trans. Audio & Electroac. AU-21: 69-79, April 1973.

[29] J.D. MARKEL & A.H. GRAY, Jr., Linear prediction of speech. Springer-Verlag, New York, 1976.

[30] J.F. MICHEL & R. WENDAHL, "Correlates of voice production". In Travis [55] (1971), 465-480.

[31] W.G. MOULTON, The sounds of English and German. (Contrastive Structure series), Univ. of Chicago Press, Chicago, 1962.

[32] R.S. NICKERSON & K.N. STEVENS, "Teaching speech to the deaf: can a computer help?". IEEE Trans. Audio & Electroac. AU-21: 445-455, Oct 1973.

[33] A.V. OPPENHEIM, R.W. SCHAFER, & T.G. STOCKAM, "Non-linear filtering of multiplied and convolved signals". Proc. IEEE 56: 1264-1291, Aug 1968.

[34] C.B. PAULSTON & M.N. BRUDER, Teaching English as a second language: techniques and procedures. Winthrop, Mass., 1976.

[35] J.M. PICKETT, "Status of speech analyzing communication aids for the deaf". IEEE Trans. Audio & Electroac. AU-20: 3-8, March 1972.

[36] R.L. POLITZER & F.N. POLITZER, Teaching English as a second language. Xerox. Mass., 1972.

[37] D.J. POVEL, "Development of a vowel corrector for the deaf", and "Evaluation of a vowel corrector as a speech training aid for the deaf". Psychol. Res. 37(1): 51-70,71-80, 1974.

[38] M.H. POWERS, "Functional disorders of articulation - symptomatology and etiology. In Travis [55] (1971), 837-875.

[39] W. PRONOVOST, "Developments in visual displays of speech information". Volta Rev, 69: 365-373, June 1967.

[40] L.R. RABINER, "On creating reference templates for speaker independent recognition of isolated words". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-26: 34-42, Feb 1978.

[41] L.R. RABINER, A.E. ROSENBERG & S.E. LEVINSON, "Considerations in dynamic time warping algorithms for discrete word recognition". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-26: 575-582, Dec 1978.

[42] L.R. RABINER et al., "Terminology in digital signal processing". IEEE Trans. Audio & Electroac. AU-20: 322-337, Dec 1972.

[43] L.R. RABINER et al., "A comparative performance study of several pitch detection algorithms". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-24: 399-418, Oct 1976.

[44] C.H. REINSCH, "Smoothing by spline functions". Numer. Math. 10: 177-183, 1967.

[45] C.H. REINSCH, "Smoothing by spline functions II". Numer. Math. 16: 451-454, 1971.

[46] H. SAKOE & S. CHIBA, "A dynamic programming approach to continuous speech recognition". In Proc. 7th Int. Cong. on Acoustics, 1971, Paper 20, p.C13.

[47] M.R. SAMBUR, "Speaker recognition using orthogonal linear prediction". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-24: 283-289, Aug 1976.

[48] M.R. SAMBUR & L.R. RABINER, "A speaker-independent digit-recognition system". Bell Sys. Tel. J. 54: 81-102, 1975.

[49] R.W. SCHAFER & L.R. RABINER, "Digital representations of speech signals". Proc. IEEE 63: 662-677, April 1975.

[50] S.R. SILVERMAN, "The education of deaf children". In Travis [55] (1971), 399-430.

[51] S.R. SILVERMAN & D.R. CALVERT, "Conservation and development of speech". In Davis & Silverman [12] (1978), 388-399.

[52] S.R. SILVERMAN, H.S. LANE, & D.R. CALVERT, "Early and elementary education". In Davis & Silverman [12] (1978), 433-482.

[53] R.E. STARK, "Teaching features of speech to deaf children by means of real-time visual displays". Proc. Int. Symp. Speech Comm. & Prof. Deafness, Washington D.C., 1972.

[54] C.K. THOMAS, An introduction to the phonetics of American English. Ronald Press Co., New York, 1958.

[55] L.E. TRAVIS (ed.), Handbook of speech pathology and audiology. Appleton-Century-Crofts, New York, 1971.

[56] N. UMEDA, "Linguistic rules for text-to-speech synthesis". Proc. IEEE 64: 443-451, April 1976.

[57] H. WAKITA, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms". IEEE Trans. Audio & Electroac. AU-21: 417-427, Oct 1973.

[58] H. WAKITA, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-27: 281-285, June 1979.

[59] G.M. WHITE & R.B. NEELY, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming". IEEE Trans. Acoust., Speech, Signal Proc. ASSP-24: 183-188, April 1976.