GLOBAL JOURNAL OF PURE AND APPLIED SCIENCES VOL. 10, NO. 3, JUNE 2004: 459 - 464
COPYRIGHT(C) BACHUDO SCIENCE CO. LTD. PRINTED IN NIGERIA. ISSN 1118-0579

AN ALTERNATIVE TOOL FOR RESEARCH IN PHONETICS: COMPUTER BASED SPEECH SIGNAL PROCESSING

E. E. WILLIAMS, R. C. OKORO and ZSOLT LIPCSEY

(Received 10 Sept. 2002; Revision Accepted 25 March, 2004)

ABSTRACT

The speech analysis and speech recognition is a topic, which is about 40 years old. However, it entered the focus of attention in the past few years again with the new ideas of instructing computers verbally, voice recognition based security systems and other areas of artificial intelligence. The languages studied are mainly non-African It is a challenge to African countries to develop these facilities for their own languages. The research in the traditional way requires quite significant funding. This paper gives support to researchers in this direction by calling their attention to the fact that using multimedia system of any modern computer can serve as a high quality research equipment for research in voice analyses. The needed part of the file format and the type of conversion needed to get the required data is discussed in the paper.

KEYWORDS: phonetics, artificial intelligence, digitiser, spectral analysis, voice processing

INTRODUCTION

Computers instructed or operated by spoken natural language instructions, voice recognition, language interpretation, understanding and automated translations are among the current research trends of Artificial Intelligence (AI) (Vonderheid, 2002; Srinivasan & Brown, 2002). The traditional equipment for studying spoken languages is usually costly and this led to the search for an alternative less expensive method. This paper therefore aims at showing that multimedia facilities of an IBM Personal Computer can facilitate high quality voice analysis suitable for research in phonetics. Therefore, the IBM Personal Computers and indeed any system which records in "Sonique" file format serves as n alternative to the traditional equipment with practically no additional cost other than the cost of a Personal Computer.

BACKGROUND CONCEPT

Natural languages are studied in both linguistics and AI (Jakobson et.al., 1963; Chomsky & Haile, 1968; Rabiner & Schafer, 1978; Russell & Norvig, 1995), with different methodologies. Both however, are split into two major methodologically different directions viz : studying a spoken language and studying written language.

Studying a spoken language requires two processes: translation of the spoken text into written text and translation of the written text into properly articulated sound waves serving as human speech (Rich & Knight, 1991; YiXu at.al. 2000; YiXu & Wang, 2000).

Research in spoken languages is the focus of this paper. Such research requires high quality recording of the human voice and this record will then be analysed and serve as the source signal sequence that is to be converted to written text. Our discussions are around this process.

Importance of the research

In Computer Science, Natural language processing has a fundamental role to play in that it forms a computational understanding of how people learn and use their native languages, and to produce a computer program that can use a human language at the same level of competence as a native language speaker.

All human knowledge has been encoded in Natural languages. Moreover, research in natural language understanding has shown that encyclopaedic knowledge is required to understand a natural language. Therefore a complete natural language using a computer system will also be a complete intelligent system (Shapiro, 1992).

Language phonetics today is the focus of attention, since the current trends in research focus among others, on systems instructed verbally. Also speech recognition is one of the tools of identification of the speaker in security procedures (Srinivasan & Brown, 2002; Gorin et. al. 2002). The results in the literature are related to English and some other overseas languages. The relevance to Africa therefore, is to explore the local languages both for translation, security and instruction. We believe that the result of our research as reported in this paper will offer a cheaper alternative tool without compromising quality to researchers in Africa.

In the first two sections of the paper we summarise the needed basics of sound and voice. In relation to voice, we will discuss basic concepts like phoneme and formation of phonemes. We will then discuss the traditional methodology of studying voice, and related basic techniques. Finally we will discuss our technique, which is a cheaper alternative to the traditional approach. In the appendix we present a voice recording based on the approach proposed.

E. E. WILLIAMS, Department of Mathematics/Statistics/Computer Science, University of Calabar, Calabar, Nigeria
R. C. OKORO, Department of Mathematics/Statistics/Computer Science, University of Calabar, Calabar, Nigeria
ZSOLT LIPCSEY, Department of Mathematics/Statistics/Computer Science, University of Calabar, Calabar, Nigeria

SOUND AND SOUND WAVES

Rabiner & Schafer (Rabiner & Schafer, 1978) gives a comprehensive summary of the basic concepts used in sound recording and analysis with special consideration to computer scientists. The discussions in this section are based on Rabiner & Schafer's ideas. More elaborate description of these topics can be found in numerous texts on classical physics and modern electronics.

When a spatial object is moving, its interaction with the air results in changes in the pressure of the air on its surface. These pressure changes travel in the air forming sound waves. They are travelling with a speed of 330mps, which is the speed of sound in the air. The moving spatial object is the source of sound. When the sound waves as pressure waves find other objects in their way, they interact with them by making the objects move as the wave pressure dictates.

The source of sound can also be a vibrating chord (which is a moving object), which may follow a simple sinusoidal periodic movement. The resulting sound wave will take the form of travelling sinusoidal waves. A vibrating chord may vibrate at several frequencies, which are determined by its length, density and its elastic properties. The human ear will hear the sound waves of various frequencies as sounds of various heights. The human ear can observe sounds from 20Hz to 20KHz. Hence studying human voice requires the recording of sinusoidal sound-wave components in this range.

The vibrating chord may be set to any starting shape and released to vibrate. Then the sound waves generated will be the superposition of sinusoidal sound waves of different frequencies with different intensities. For any sound its composition from sinusoidal waves of different frequencies is unique and knowing the frequency components of a sound and their amplitudes, one can reproduce the sound. To determine the frequency composition of a sound wave is the topic of the spectral analysis (Koenig et.al. 1946). Spectral analysis is one of the basic techniques to study sound waves.

When we observe or record sound, we use a receptor (ear, microphone, chord, membrane, etc.). We assume that we are at a fixed point of the space, and we are recording the sound at that fixed location. Fixed location may not be so important since one can record sound from a moving object as well. What is important is that our observations are made by a "small" size equipment, and the observed data is from a given location of the space.

Recording sound waves

The classical method of recording sound is analogue recording, which means that the pressure intensity at the point of observation is converted proportionally to voltage intensity, magnetic field intensity on a tape, and it is reproduced as changing pressure wave when it is processed and sent to the loud speaker.

The problems with this process come from the noise. The original signal may be noisy, and it is difficult to correct it. The amplifier may add its own noise, and that leads to another source of noise. Finally there is a limit in the fidelity of such system.

The modern technology developed is the digital recording technique where the analogue signal is sampled and the sampled digital information is recorded and used (Shannon 1948; Rabiner & Levinson 1981; Padmanabhan & Picheny, 2002).

The sampling process is called digitisation. It takes the following procedure:

The amplitude of the analogue signal is determined at regular time points. The number of amplitude readings in a second gives the sampling frequency. The range of amplitudes is partitioned into 256 classes (may be 65535 classes), and the class, which the observed amplitude belongs to, is recorded (Figure 1.). Hence we record nonnegative numbers between 0 and 255 (or between 0 and 65535 accordingly).

The process of digitisation is controlled by the following theorem (Shannon, 1948):

Theorem (Shannon): If the highest frequency to consider in the sound wave is fr then the sampling frequency must be at least 2fr.

In other words, the digitised records give accounts of frequency components with frequency ≤0.5×sampling frequency. To fully analyse speech signals, one may need a sampling frequency of at least 40,000Hz.

FORMATION OF SPEECH AND THE TARGET OF VOICE ANALYSIS

When voice is mentioned, we are referring to a sound, produced by human beings, in this context, while they are speaking. The mouth, throat, tongue etc., from the vocal cords or glottis to the lips is called the vocal tract and it forms a vibrating system. Their physical properties are changed during speaking and these changes will alter the frequency composition of the sound produced, and we interpret it as speech (Rabiner & Schafer, 1978).

The effect of sound generation and propagation is similar to the resonance effect observed in organ pipes and wind instruments. In the context of speech production, the resonance frequencies of the vocal tract tube are called formant frequencies and depend on the shape and dimensions of the vocal tract. The spectral properties of the speech signal vary with time as the vocal tract shape varies. The time varying spectral characteristics of the speech signal can be graphically displayed through the use of the sound spectrogram (Flanagan, 1972; Keonig et al, 1946).

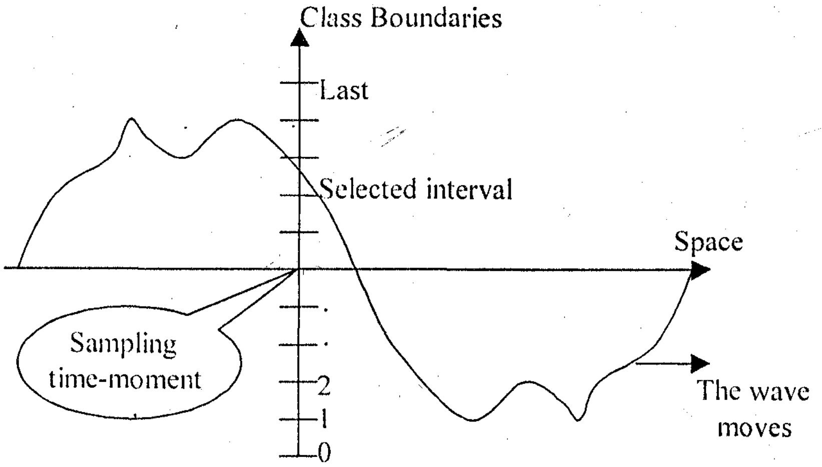

Figure 1: Sampling process: The wave moves across the sampling point and in each sampling moment the amplitude is sampled.

What sounds and how they are generated depends on the culture where the person grew up and the language the person speaks. Different languages use different sets of sounds for speech. In addition, different dialects of the same language may use different sets of sounds. We will discuss the sets of sounds used in the next section.

Phonetics and Phonology

As discussed in the introduction, language processing can be divided into two main tasks.

1. Processing written text;
2. Processing spoken language using all the information in 1 in addition to the knowledge about phonology as well as enough added information to handle the ambiguities that arise in speech (Rich and Knight, 1991).

Phonetics: The study of the sounds used in human speech is the subject of phonetics. It concerns the sounds of human speech and providing a symbol for each sound element (called phoneme) of the language. Speech is perceived as a string of groups of phonemes. Every language has a small set (between 20 and 50) of phonemes. English for example, can be represented by a set of 42 phonemes (Rabiner & Schafer, 1978), some of which are unique to it, while some occur in other languages also.

All the sounds that occur in other languages of the world have been classified into two groups: Vowels and Consonants. Each vowel or consonant has been assigned a symbol called phonetic symbol. The International Phonetic Association has agreed upon these symbols.

Branches of phonetics

The study of phonetics has been classified into three main categories as follows:

1. Articulatory phonetics: Seeks to explain and classify speech sounds in terms of the organs that produce them and how these organs function to produce the speech sounds
2. Acoustic phonetics: Sees speech in terms of the properties of sound waves generated when the speech organs go into action, and how the waves travel from speaker to the hearer and interpreted as sound.
3. Auditory Phonetics: Sees sounds in terms of how they are perceived and interpreted by the hearer.

This paper focuses on (2). To study the sound waves generated by human being requires their recording, and analysis. Conventionally the processes and equipment involved in the recording and analysis are as follows:

Microphone: This is a transducer, which changes its resistance or capacitance or generates voltage variation according to the pressure of the sound waves based on inductivity. The input is sound, the output is varying voltage which is proportional to the variation of the pressure in the sound wave. The microphone needed for this type of HI.FI. recording must be a microphone with high acoustic impedance like the capacitor type microphone.

Digitiser: The digitiser observes the signal voltage at regular intervals, the frequency of which is at least twice the considered highest frequency component of the signal. The voltage range is partitioned into equal intervals, numerated from 0 to the number of classes (256 normally). The digitiser determines, which partition contains the observed voltage, and records the sequence number of the interval. The situation is shown on Figure 1.

Oscilloscope/Spectral analyser: The

oscilloscope plots the waveform in an interval determined by the frequency of the oscilloscope (adjustable). If the waveform is a regular shap.ɜ and frequency of the oscilloscope matches the wave frequency then one can achieve that the dynamic plot is stand still. This is a tool suitable for determining the frequency of the wave.

ALTERNATIVE SOLUTION USING MULTIMEDIA

Our approach simply is to use any recording system, which records in 'Sonique' file format, which is a digitised record. One of such recording systems, which can be found easily is the built-in multimedia facilities of an IBM PC. Since the analysis requires a computer, using it is economical.

We assume that an IBM PC is available, with multimedia sound system activated. The system has a High-Fidelity (HI-FI) microphone and HI-FI loud speakers as accessories.

We then can record any sound by using the Sound Recorder facility of WINDOWS 95, 98, 2000 or any other version of WINDOWS that has the facility.

The result will be a Sonique file, which has a file header of 60 bytes, and this is followed by the sound data, which is a sequence of bytes for mono record, a sequence of 2 times two bytes unsigned integers for stereo records. Each byte for mono specifies one out of the 256 classes, and each two bytes of stereo specifies one out of 65536 classes.

To enable us use this file for the purposes of the above analysis, we convert the data segment of the file to a tabulated ASCII file so that if 'z_j' is the current data item (class number), then we convert it to decimal signed number as

$z_j \rightarrow z_j - 128$ where z_j refers to one of the 256 classes;

$z_j \rightarrow z_j - 32768$ where z_j refers to one of the 65536 classes for stereo.

Since the data is now in text file format, it is portable to most of the application packages. A C++ program written for the purpose does the conversion. Table 1 in Appendix illustrates a sample-converted data using a recorded vowel. We used MS-EXCEL to plot selected segments of the wave to obtain the needed waveforms as indicated in Figure 2.

TECHNICAL SPECIFICATIONS/RECORDING QUALITY

Multimedia components

1. There must be an IBM compatible computer system preferably the Pentium series for better quality.
2. There must be a sound card installed in the system, it could be on board or external board. This must be connected to the computer expansion slot for the external card or plugged into the sound connector on the motherboard of the system.
3. The required device driver software must be installed to make sure that the device gets the parameters it requires to function well
4. There should be a high quality microphone preferably with high acoustic impedance. It should be small in size to avoid diffraction. The frequency response should be flat i.e. within the audible range of 20Hz to 20KHz.
5. A loudspeaker for the output of the recorded sound
6. A CD-ROM drive for the device software installation.
7. CD-ROM to sound card cable to connect the sound from CD-ROM drive to the sound card

SAMPLING FREQUENCIES

QUALITY	ATTRIBUTES	FORMAT
CD-QUALITY	44,100Hz, 16bit stereo 176KBps	PCM
RADIO QUALITY	22,050Hz, 8bit mono 22KBps	PCM
TELEPHONE QUALITY	11,025Hz, mono 11KBps	PCM

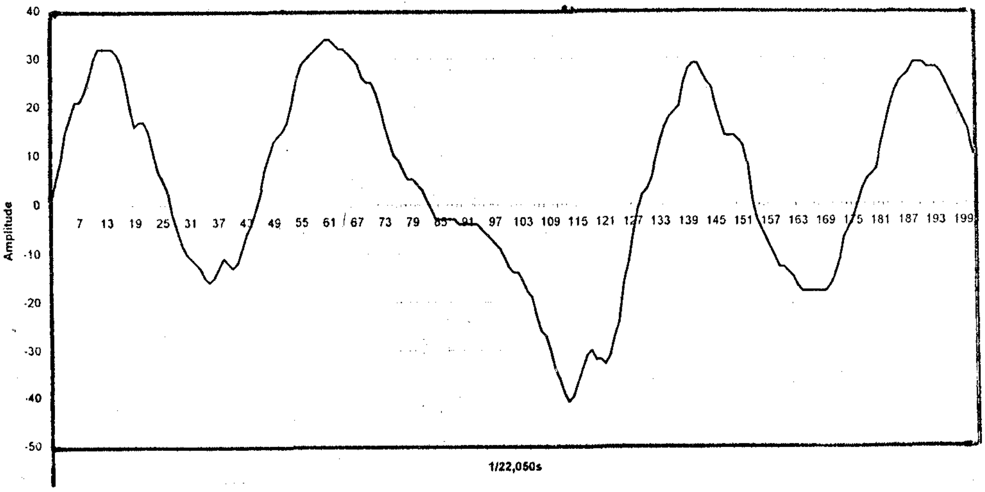

Figure 2: Plotting the sampled voice data presented in table 1 (A selected section of 200 samples). The plotted curve covers about 1/100 second period of time.

There are many other formats and attributes but the most widely used are those shown above for recording.

CONCLUSION

Instruction to computers and speech analysis is mostly done in English and in some non-African languages. The available facilities in those languages are provided by those nations such as Japan, India, China etc.. It is a challenge for Africans to open avenues to their own people in the use of modern facilities of information processing. This paper calls the attention to a possible approach to carrying out the necessary research with minimal investment. The inbuilt multimedia facilities of an IBM PC provide high quality speech processing, which otherwise could have been quite expensive. We believe that this can contribute to the developments in the field of speech studies.

REFERENCES

Chomsky, N. and Haile, M., 1968. The sound pattern in English, Harper and Row Publishers.

Flanagan, J. L., 1972. Speech analysis, synthesis and perception. Second Ed Springer-Verlag.

Gorin, A.L., Abella A., Alonso T., Riccardi G. and Wright J. H., 2002. Automated Natural Spoken Dialog. Computer IEEE, 35 (4): 51-56.

Jakobson, R., Fant, C. G. M. and Haile, M., 1963. Preliminaries to speech analysis: The distinctive features of their correlates, MIT Press.

Keonig W, Dunn H. K. and Lacy L.Y., 1946. The sound spectrograph J. Acoust. Soc. Am. 17: 19-49.

Padmanabhan, M. and Picheny, M., 2002. Large vocabulary speech recognition algorithms. Computer IEEE, 25(4): 42 - 50.

Rabiner, L. R. and Levinson S., 1981. Isolated and connected word recognition – Theory and selected applications. IEEE Transactions on Communications, 29(5): 621-659.

Rabiner, L. R., and Schafer, R.W., 1978. Digital Processing of speech signals, Prentice Hall.

Rich, E. and Knight, K. 1991. Artificial Intelligence 2nd Ed.

Russel, S. J. and Norvig P., 1995. Artificial Intelligence. A modern approach. Prentice Hall.

Shannon, C. E., 1948. A mathematical Theory of Communication. Bell Systems Technical Journal, 27: 379-423.

Shapiro, Stuart, Encyclopaedia of AI, Vol. 1, 1992.

Srinivasan, S. and Brown, E., 2002. Is speech recognition becoming mainstream? Computer, IEEE, April.

Stephen J. Bigelow, 2000. Troubleshooting, Maintaining and Repairing PCs Millennium Edition.

System Board Resources, 2000. A handbook on system board installation for socket 7 processor PC133.

Vonderheid, E., 2002. Speech Synthesis Offers Realism for 'Voices' of Computers, Automobiles and Yes, Even VCRs. The Institute, IEEE, March.

Winograd, T., 1983. Language as a cognitive process Vol.1 Addison Wesley.

YiXu, Liberman, A. M. and Whalen, D. H., 2000 On the Immediacy of Phonetic Perception, Psychological Science, 8: 358 – 362.

YiXu and Wang E., 2000. Phonetic Targets as a link between Speech Production and Speech perception. Journal of the Acoustical Society of America, 108: 2531 – 2532.

APPENDIX.

Table 1. The following sampled wave data is taken from a digitised record. 200 samples cover about 1/100 second.

S/N	Wave	S/N	Wave	S/N	Wave	S/N	Wave	S/N	Wave
1	1	41	-12	81	3	121	-33	161	-14
2	5	42	-9	82	1	122	-31	162	-15
3	9	43	-6	83	-1	123	-27	163	-17
4	15	44	-4	84	-3	124	-24	164	-18
5	18	45	-1	85	-3	125	-16	165	-18
6	21	46	2	86	-3	126	-12	166	-18
7	21	47	7	87	-3	127	-6	167	-18
8	23	48	10	88	-3	128	-1	168	-18
9	26	49	13	89	-4	129	2	169	-18
10	30	50	14	90	-4	130	3	170	-17
11	32	51	15	91	-4	131	5	171	-15
12	32	52	17	92	-4	132	9	172	-12
13	32	53	21	93	-4	133	13	173	-7
14	32	54	26	94	-5	134	16	174	-5
15	31	55	29	95	-6	135	18	175	-3
16	29	56	30	96	-7	136	19	176	0
17	25	57	31	97	-8	137	20	177	3
18	20	58	32	98	-9	138	25	178	5
19	16	59	33	99	-11	139	28	179	6
20	17	60	34	100	-13	140	29	180	7
21	17	61	34	101	-14	141	29	181	12
22	15	62	33	102	-14	142	27	182	16
23	11	63	32	103	-16	143	25	183	20
24	7	64	32	104	-18	144	24	184	23
25	5	65	31	105	-19	145	20	185	25
26	3	66	30	106	-23	146	17	186	26
27	-2	67	29	107	-26	147	14	187	27
28	-5	68	26	108	-27	148	14	188	29
29	-8	69	25	109	-30	149	14	189	29
30	-10	70	25	110	-34	150	13	190	29
31	-11	71	23	111	-36	151	12	191	28
32	-12	72	20	112	-39	152	8	192	28
33	-13	73	16	113	-41	153	1	193	28
34	-15	74	13	114	-40	154	-3	194	27
35	-16	75	10	115	-37	155	-5	195	25
36	-15	76	9	116	-34	156	-7	196	23
37	-13	77	7	117	-31	157	-9	197	21
38	-11	78	5	118	-30	158	-11	198	19
39	-12	79	5	119	-32	159	-13	199	17
40	-13	80	4	120	-32	160	-13	200	15