DEVE LOPING A SPEECH PRODUCTION MATHEMATICAL MODEL IN TERMS OF THE Z-TRANSFORM FOR SPEECH RECOGNITON IMPLEMENTATION IN THE COMPUTER

Majed Ismail Hussein

Department of MIS, Al- Isra University
Email: drmajed1@hotmail.com

ABSTRACT

The use of speech recognition systems allow messages to interact with the computer in natural it is possible to solve Voice control of the computing process(such as speech input commands WINDOWS), Voice control of various technical systems, Automatic entry of textual information in the personal computer to fill in the various documents databases. The use of speech recognition providers Increasing management efficiency through the use of verbal interaction channel (average specialist computer systems from the keyboard can enter only 10-20 words / min ,and the voice can convey 100-200 words /min) A mathematical model of speech production can be described in terms of Z-transform in order to implement it on a computer in the form of the relationship.

Keywords: *speech recognition, speech production, speech signal, mathematical model, z-formation*

1. INTRODUCTION

Speech recognition systems. can be classified on Automatic speech recognition - the process by which a computer displays the acoustic speech signal into text or corresponding commands that control the computing process. A more complex notion is automatic speech recognition which includes automatic recognition and semantic analysis of the recognized text. Recent systems are currently in the stage of research development Speech recognition system can be divided can be classified on several grounds, including : Depending on the speaker, The volume of the dictionary The nature of the recognized speech flow.

Consider these characteristics. Speaker-dependent systems are designed for single users. These systems are usually simpler in structure, it is cheaper but lack the flexibility and the ability to adapt to a group of speakers or to work with unknown Announcer.

Announcer - independent systems are designed to operate with any particular type of speaker. These systems are the most difficult to develop, are the most expensive and provide the recognition accuracy is lower than the speaker-dependent systems. However, they are more flexible and easy to use.
System with adaptation for speaker allows configuration for a particular speaker or group of speakers. Such systems can provide enough efficiency of recognition, but the process of setting it under a speaker or group of speakers can be quite time consuming. The size of vocabulary speech recognition system is directly related to its complexity and strongly influences the characteristics of recognition accuracy. Dictionary size is determined by the specific requirements of the relevant application system. Some applications only require a few words (for example, only numbers), others require very large dictionaries (eg dictation system for automatic text) Usually covers the following grades of dictionaries A small vocabulary - tens of words Average vocabulary - hundreds of words The Big Dictionary - a thousand words A very large vocabulary - tens of thousands of words. isolated words and continuous speech Systems of recognition of isolated words are focused on recognition of single words at the same time demand a pause between saying each word. This - the simplest form of recognition, because there are easier to find the final and initial points of the words and pronunciation with neighboring words do not a By the nature of the speech stream speech recognition system can be divided into recognition of ffect each other, which provides enough high-quality recognition.

2. BASIC CONCEPTS

2.1 Speech production

Speech production is the process by which spoken words are selected to be produced, have their phonetics formulated and then finally are articulated by the motor system in the vocal apparatus. Speech production can be spontaneous such as when a person creates the words of a conversation, reaction such as when they name a picture or read aloud a written word, or a vocal imitation such as in speech repetition.[12,13] Speech production is not the same as language production since language can also be produced manually by signs.

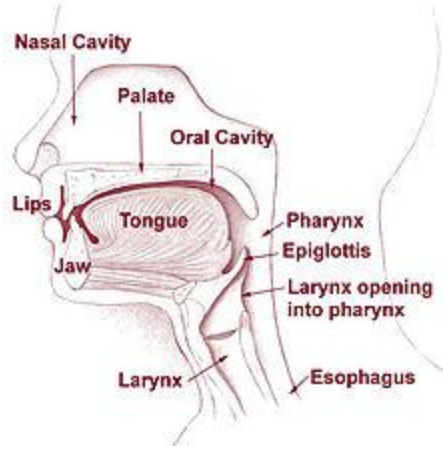

fig.1 Human vocal apparatus used to produce speech

In ordinary fluent conversation people pronounce each second roughly four syllables, ten or twelve phonemes and two to three words out of a vocabulary that can contain 10 to 100 thousand words.[1] Errors in speech production are relatively rare occurring at a rate of about once in every 900 words in spontaneous speech.[2] Words that are commonly spoken or learned early in life or easily imagined are quicker to say than ones that are rarely said, learnt later in life or abstract.[3,4] Normally speech is created with pulmonary pressure provided by the lungs that generates sound by phonation in the glottis in the larynx that then is modified by the vocal tract into different vowels and consonants. However speech production can occur without the use of the lungs and glottis in alaryngeal speech by using the upper parts of the vocal trait. An example of such alaryngeal speech is Donald Duck talk[5] The vocal production of speech can be associated with the production of synchronized hand gestures that act to enhance the comprehensibility of what is being said[6] The production of spoken language involves three major levels of processing.[1,7,8] The first is the processes of conceptualization in which the intention to create speech links a desired concept to a particular spoken word to be expressed. Here the preverbal intended messages are formulated that specify the concepts to be verbally expressed. This is a competitive process in which an appropriate word is selected among a cohort of candidates.[1,7,8] The second stage is formulation in which the linguistic form required for that word's expression is created. This process involves such processes as the generation of a syntactic frame, and phonological encoding which specifies the phonetic form of the intended utterance. At this stage a lemma is picked that is the abstract form of a word that lacks any information about the sounds in it (and thus before the word can be pronounced). It contains information concerning only meaning and the relation of this word to others in the sentence.[1,7,8] The third stage is articulation which involves the retrieval of the particular motor phonetics of a word and the motor coordination of appropriate phonation and articulation by the lungs, glottis, larynx, tongue, lips, jaw, and other parts of the vocal apparatus[7] Speech production motor control in right handers depends mostly upon areas in the left cerebral hemisphere. These areas include the bilateral supplementary motor area, the left posterior inferior frontal gyrus, the left insula, the left Primary motor cortex and temporal cortex[9] There are also subcortical areas involved such as the basal ganglia and cerebellum.[10,11] The cerebellum aids the sequencing of speech syllables into fast, smooth and rhythmically organized words and longer utterances.[11]

2.2 A linear mathematical model of speech production.

The linear model of speech production was developed in the late 50s, its mathematical justification and a detailed study carried out in [3,7] on the basis of carefully posed experimental studies. The corresponding block diagram is shown in fig.2

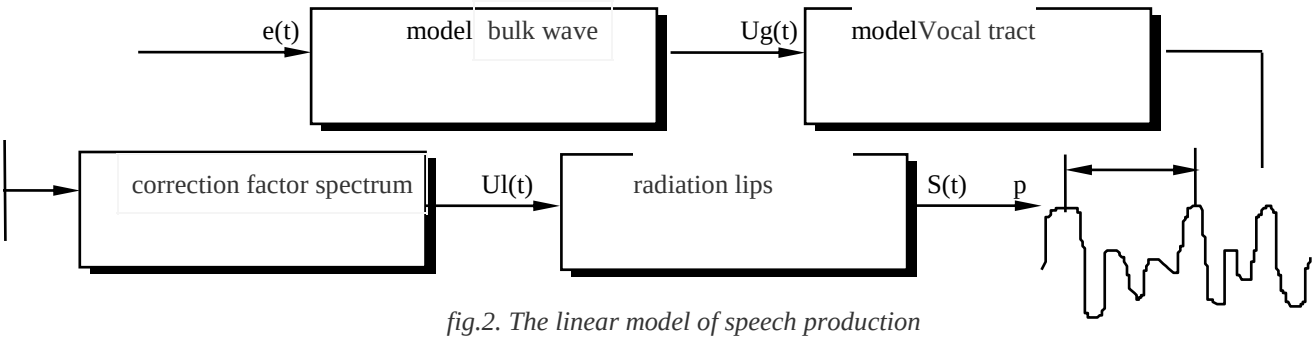

fig.2. The linear model of speech production

The bulk wave velocity in the glottis, Ug (t) is modeled by a bipolar output lowpass filter with a cutoff frequency of 100 Hz. Filter input signal e (t) is a pulse sequence with period P for vowelize sounds and random noise with a uniform spectrum for nevokalizovannyh sounds. It should be noted that this model is a special case of a more general model, since it is not mixing pulse and noise signals for simulation vowelize fricatives or connect another branch with a filter to simulate the nasal sounds. Vocal tract in this model is a pole filter consisting of a small group of bipolar cascade connection of resonators. Each resonance is defined here as the relevant formant frequency and bandwidth.

A more accurate model requires further include an infinite number of resonators, which should provide in the main rise of the spectrum at low frequencies. Therefore, when you need to accurately simulate the characteristics of the speech production only at low frequencies, for example, the most important part of the range of low frequencies from 20 Hz to several kHz, then this form of the spectrum can be obtained using a correction factor to account for the impact on low-frequency spectrum All poles, far more common on the frequency axis, almost regardless of their settings. The bulk velocity of the wave lip Ul (t) is converted into acoustic vibrations of air at some distance from the mouth (these are the fluctuations and represent speech wave S (t) using a model of radiation lips).

3. METHODOLOGY

A mathematical model of speech production in terms of the z-transform [7] in order to implement it on a computer in the form of the relationship .

$$S(z) = E(z)\,G(z)\,V(z)\,L(z), \tag{1}$$

where $S(z) \longleftrightarrow s(nt)=s(t)\Big|_{t=nt}$ (2)

which indicates the correspondence between the continuous signal s (t), its discrete copy of s (nT), obtained by sampling s (t) with an interval of T, and z-transform S (z). Usually for short interval considered T = 1, so that s (n) describes the result of sampling s (t). For other variables the interval T is also assumed to be normalized. Excitation signal at the input of the model described by the function of the glottis

E (z) ßà e (n) and a sequence of readings of unit amplitude with repetition period equal to the period of the fundamental tone P = IT, where

I - a positive integer, ie

$$E(z) = \sigma\sum_{n=0}^{\infty} \left(z^{-1}\right)^n = \frac{\sigma}{1-z^{-1}} \tag{3}$$

for | z |> 1. The transfer function of the glottis G (z) has the form

$$G(z) = \frac{1}{\left(1-e^{-cT}z^{-1}\right)^2} \tag{4}$$

a transfer function model of radiation lip L

$$L(z) = 1 - z^{-1}. \tag{5}$$

All these simplifying assumptions do not allow using a linear model to predict the concrete realization of the speech process.

Pole transfer function model of the vocal tract V (z), products containing formant has the form

$$V(z) = \left\{\prod_{t=1}^{K}[1 - 2e^{-c_i T}\cos(b_i T)z^{-1} + e^{-2c_i T}z^{-2}]\right\}^{-1} \tag{6}$$

where the frequency and the bandwidth of the i-th formant are evaluated according to the formulas and. The digital representation of this model, the correction term that takes into account the effect of the poles at higher frequencies, can be excluded.

For z = 0 zeros do not affect the determination of transfer functions that contain only poles or only zeros. For example, the function G (z) can be written in two equivalent forms:

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2} = \frac{z^2}{(z - e^{-cT})^2} \quad (7)$$

In other words, if z = 0, the poles and zeros are usually not taken into account when calculating the total number of poles and zeros. A model of speech production requires submitting to its input only a quasi-periodic pulse train or a random sequence and fully characterized by a set of frequencies and bands of formants. Thus determined only by the vowels and fricative sounds in the steady state. However, this model is easy to implement and arbitrary input signal e (n) and parameters of the function V (z), which are changed or adjusted at the required intervals to represent time-varying nature of the speech signal.

Mandatory procedure to be used in speech synthesis, is the restructuring of the model parameters from a sequence of speech formation path e (n) at the beginning of each period of the fundamental tone (called fusion, with simultaneous c frequency of the fundamental tone).
Combining the transfer functions of models of the glottis G (z), the vocal tract V (z) and lip radiation L (z) has the form

$$G(z)V(z)R(z) = \frac{(1 - z^{-1})}{(1 - e^{-cT} z^{-1})^2 \{\prod_{i=1}^{K} [1 - 2e^{-c_i T} \cos(b_i T) z^{-1} + e^{-2c_i T} z^{-2}]\}} \quad (8)$$

where K - number of formants, defining the model.

Numerator 1-z-1 is almost reduced to one of the factors in the denominator [1-exp (-cT) z-1], since the exponent significantly less than 1 kT. This discrete model can be simplified speech synthesis, and further, making it a pole, ie

S (z) = E (z) A (z) (model synthesizer), (9)

where A (z) is defined as follows:

$$A(z) = \sum_{i=0}^{M} a_i z^{-i} (a_0 = 1) \approx \frac{1}{G(z)V(z)L(z)} \quad (10)$$

when M> = 2k +1.

Filter with transfer function A (z) contains only zeros and then it will be called the inverse filter. Filter with transfer function 1 / A (z) - a pole filter, which allows us to describe the behavior of the smoothed spectrum of the speech signal with a constant factor.

Equation (8) is a mathematical model of synthesis, because if the signal, z-transform is equal to E (z), is fed to the pole filter with characteristic 1 / A (z), then its output and is a model of the speech signal, z transform which is denoted by S (z) '. Multiply both sides of (8) for A (z) provides a model for analysis of the speech signal:

E (z) = S (z) A (z). (11)

4. CONCLUSIONS AND DISCUSSIONS

This equation is a mathematical model of analysis of speech, as if the voice signal S (z) is input to the inverse filter with the characteristic A (z) (whose coefficients are determined by analyzing the speech signal), the output will be E (z)-excitation function speech signal.

The parameters that define the model of speech production or synthesis, are the coefficients ai, i = 1,2 ,..., M filter characteristic 1 / A (z) and parameters of the function E (z) - the period of the main tone and gain of P s .

5. REFERENCES

[1]. Levelt, WJ "Models of word production.". *Trends in Cognitive Sciences* 3 (6): 223–232. doi:10.1016/S1364-6613(99)01319-4. PMID 10354575. (1999). http://www.columbia.edu/~rmk7/HC/HC_Readings/Levelt.pdf.

[2]. Garnham, A, Shillcock RC, Brown GDA, Mill AID, Culter A "Slips of the tongue in the London–Lund corpus of spontaneous conversation". *Linguistics* 19 (7–8): 805–817. doi:10.1515/ling.1981.19.7-8.805. http://dare.ubn.kun.nl/bitstream/2066/15615/1/6017.pdf.(1981).

[3]. Oldfield RC, Wingfield A ("Response latencies in naming objects". *Quarterly Journal of Experimental Psychology* 17 (4): 273–281. doi:10.1080/17470216508416445. PMID 5852918.(1965).

[4]. Bird, H; Franklin, S; Howard, D "Age of acquisition and imageability ratings for a large set of words, including verbs and function words". *Behavior Research Methods, Instruments, & Computers* 33 (1): 73–9. doi:10.3758/BF03195349. PMID 11296722. http://brm.psychonomic-journals.org/content/33/1/73.full.pdf.(2001).

[5]. Weinberg, B; Westerhouse,. "A study of buccal speech". *Journal of Speech and Hearing Research* 14 (3): 652–8. PMID 5163900.J (1971)
[6]. McNeill D. *Gesture and Thought*. University of Chicago Press. ISBN 978-0-226-51463-5.(2005)
[7]. Levelt, WJM. *Speaking: From Intention to Articulation*. MIT Press. ISBN 978-0-262-62089-5.(1989)
[8]. Jescheniak, JD; Levelt, WJM \. "Word frequency effects in speech production: retrieval of syntactic information and of phonological form". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (4): 824–843. doi:10.1037/0278-7393.20.4.824. CiteSeerX: 10.1.1.133.3919.(1994)
[9]. Indefrey, P; Levelt, WJ. "The spatial and temporal signatures of word production components". *Cognition* 92 (1–2): 101–44. doi:10.1016/j.cognition.2002.06.001. PMID 15037128.(2004)
[10]. Booth, JR; Wood, L; Lu, D; Houk, JC; Bitan, T "The role of the basal ganglia and cerebellum in language processing". *Brain Research* 1133 (1): 136–44. doi:10.1016/j.brainres.2006.11.074. PMC 2424405. PMID 17189619. //www.ncbi.nlm.nih.gov/pmc/articles/PMC2424405/.(2007).
[11]. Ackermann, H. "Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives". *Trends in Neurosciences* 31 (6): 265–72. doi:10.1016/j.tins.2008.02.011. PMID 18471906(2008)
[12]. Hickok, G.. "The role of mirror neurons in speech and language processing". *Brain and Language* 112 (1): 1–2. doi:10.1016/j.bandl.2009.10.006. PMC 2813993. PMID 19948355. //www.ncbi.nlm.nih.gov/pmc/articles/PMC2813993/.(2010)
[13]. Skoyles, J. R. "Mapping of heard speech into articulation information and speech acquisition". *Proceedings of the National Academy of Sciences* 107 (18): E73. doi:10.1073/pnas.1003007107. . (2010)