To The Problem of Speech Recognition

Kudubayeva S. A. - Candidate of Technical Sciences, Docent
Akhmet Baytursynov Kostanay State University
Kazakhstan, Kostanay

Ermagambetova G. N. - M.Sc.
Akhmet Baytursynov Kostanay State University
Kazakhstan, Kostanay

Abstract

In the given article peculiarities of the Kazakh speech recognition as well as algorithms to create a recognition component for speech commands which are to be used to design applications are considered. A short survey of digital signals processing techniques, dynamic programming methods is carried out; theoretical aspects of algorithms are provided.

Introduction

From the physical point of view speech consists of a sequence of speech sounds with pauses between their groups [1]. At a normal speech pace pauses appear between fragments of phrases, because in such a case words are pronounced together (although ear, as a rule, percepts words separately). At a slow pace, for example while dictating, pauses can be made between words and even between their parts. One and the same speech sound is pronounced in different waysby different people, because every person has his own manner of speech sounds' pronunciation. Pronunciation of speech sounds depends upon stress, neighboring sounds, etc. But with all the diversity in their pronunciation they represent physical realizations (pronunciation) of a limited number of generalized speech sounds which are called phonemes. Phoneme is what a person wants to pronounce and speech sound is what a person actually pronounces. Phoneme in its relation to speech sound plays the same role as the standard letter doesin relation to its handwritten form in a particular writing.

There are 40 main and 2 indefinite phonemes in the Kazakh language.

Speech sounds are divided into voiced and voiceless and into sonorant, hard and soft sounds. Voiced sounds are produced with the help of vocal cords being in this case in the stressed state. Under the pressure of the air coming from the lungs, they periodically move apart, resulting in a discontinuous flow of air. The pulses of air flow generated by the vocal cords with sufficient accuracy can be considered to be periodic. The corresponding pulse repetition's period is called the period of the main tone of voice$T0$. Aninversevalue T0,i.e.$1/T0$, is called pitch frequency. If vocal cords are thin and very tense, the period is short and the pitch frequency is high; if vocal cords are thick and lax, the pitch frequency is low. Pitch frequency for all the voices lies in the range 70...450 Hz. While making a speech the pitch frequency changes in accordance with the stress and accentuation of sounds and words as well as to show emotions (question, exclamation, surprise, etc.). Pitch frequency's change is called intonation. Every person has his own range of changing the main tone (usually it is a little bit more than an octave) and his own intonation. The latter is of utmost importance to recognize the speaker. The main tone, intonation, oral handwriting and voice timbre are used to identify a person and the degree of identification's reliability is higher than while using the fingerprints. The main tone's pulses have a sawtooth shape, and because of that during their periodical repetition we get a discrete spectrum with a large number of harmonics (till 40), the frequencies of which are multiples of that of the main tone's. Spectrum envelope of the main tone's pitch tends to decline to higher frequencies with a slope of about 6 dB / oct. That is why for a man's voice the constituencies' level is about 3000 Hz what is below their level of 100 Hz for about 30 dB. While pronouncing voiceless sounds, vocal cords are lax, the air stream from the lungs comes easily to the mouth cavity. Meeting on its way a variety of obstacles in the form of a tongue, teeth, lips, it forms a vortex, creating noise with a continuous spectrum.

According to the noise and voice participation consonants are divided into nasal sonorants (м,н,ң), smooth sonorants (р,й,у); voiced (б, в, д, г, ж, з, й) and voiceless (п, ф, т, с, ш, к, қ, х, ц, ч) sounds. According to the tongue position vowel phonemes are divided into hard (а,о, ұ,ы,у) and soft (ә, ө, і, е, ү, и).

According to the formation's place phonemes are divided into labial, non-labial, broad and narrow, nasal, labio-dental, forelingual, mediolingual, backlingual (velar), guttural, stops, fricatives, vibrating consonants. There are 9 specific sounds in the Kazakh language as against the Russian language. They are: ә – soft vowel, ғ – guttural consonant, қ- voiceless backlingual consonant, ң –nasal backlingual stop, ө- broad labial softvowel, ұ- hard narrow labial vowel, ү–soft variant of ұ-sound, h- guttural consonant sound, ы- hard vowel, і– soft non-labial narrowvowel.

While pronouncing speech sounds a tongue, lips, teeth, a lower jaw, vocal cords should be in a strictly defined position or movementfor each separate phoneme. These movements are called articulationof speech organs. In this case in the speech formation tract resonant cavities defined for the given phoneme are created and for the continuous sounding of speech phonemes some transitions from one tract's form to another one are produced.

In the process of speech sounds' pronunciation either tonal pulse signal or noise signal or both of them go through the speech path. Speech path is a rather complicated acoustic filter with a number of resonances created by mouth, nasal and nasopharyngal cavities, i.e. with the help of organs of speecharticulation. Consequently, even tonal or noise spectrum is transformed into a spectrum with a number of maxima and minima. Maxima of a spectrum are called formants and zero failures are called antiformants. For each separate phoneme spectrum envelope has its individual and rather definite form. While making a speech its spectrum constantly changes and formant transitions are produced.

Frequency range of speech is between 70...7000 Hz. Voiced sounds, especially vowels, are characterized by a high level of intensity, voiceless sounds have a low level. While making a speech its loudness constantly changes. It changes drastically when pronouncing plosives. Dynamic range of speech levels is within 35...45 dB. Vowels have an average duration of about 0.15 s., consonants – about 0.08 s. (**п**sound-about 30 ms.).

Speech sounds are differently informative. Thus, vowels contain not so much information about the meaning of speech, whereas voiceless consonants are more informative (for example, in the word "қорғау" a sequence of "о, а, у" does not have any sense, while "қ, р, ғ" provides for an answer about the meaning). Hence, speech intelligibility is reduced under the influence of noise, primarily due to the masking of voiceless sounds.

It is known that in order to transmit the same message by telegraph and through the speech path different bandwidth is needed. A telegraph message is sufficient with the bandwidth of no more than 100 bit/s., a speech message lasts for about 100000 bit/s. (the band is equal to 7000 Hz, dynamic range is 42 dB, i.e.seven-digit code is necessary, hence we have 2·7000·7=98000 bit/s.), what is 100 times more.

Speech sounds' formation occurs by giving commands to the muscles of the organs of speech articulation from the brain's speech center. The total flow of messages from it is in average no more than 100 bit/s. All other information in the speech signal is called accompanying information.

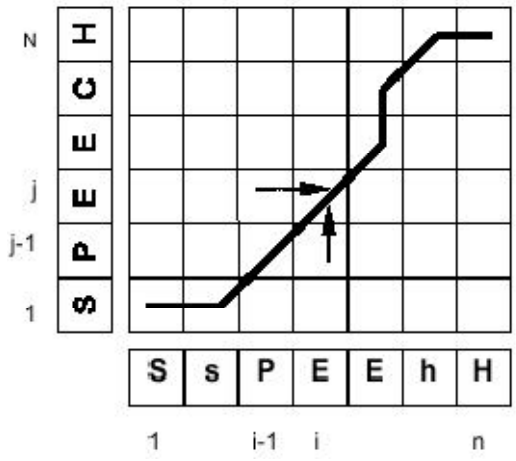
Speech signal represents a kind of a modulated carrier unit. Its spectrum is $p(w) = E(w) * F(w)$, where$E(w)$ is the generating function's spectrum, i. e. spectrum of the main tone or noise's pulses; $F(w)$is the filtering function of the speech path which is called modulating curve. This modulation refers to the group of special spectral ones. In such a case a carrier unit has a broadband spectrum resulting in modulation changing the ratio between the frequency components, i.e. the form of spectrum envelope changes. Nearly all the information about speech sounds is contained in the spectrum envelope of speech and its change in time (information about speech sounds is partially kept in the transitions from tonal spectrum to the noise one and from noise spectrum to the tonal one; by such transitions it is possible to know about the replacements of voiced sounds by voiceless ones and vice versa). All these changes occur slowly (at a speech pace).

To render the meaning of speech it is enough to render information about the form of spectrum envelope of speech and its change in time at a pace of speech sounds' change as well as information about the change of the main tone of speech and of tone-noise groups.

The above regularities of speech making form a complicated multifrequency signal, which is to be processed to extract information. For these purpose different kinds of transformations, e.g. Furye are used.

Speech is a process which changes with time. Different ways of pronouncing one and the same word usually have different duration, and the pronunciation of one and the same word with the equal duration differs in the middle part because of different parts of a word, which are pronounced with different speed.

In order to get the global mark for differences between the two speech samples represented as vectors alignment in time, as it is shown on Picture 1, should be made. On the given picture the model is shown vertically and the incoming signal – horizontally. On the picture an incoming signal "SsPEEhH" is a "noisy" version of the "SPEECH" model. The idea of an algorithm of the dynamic time distortion (DTD) lies in the fact that "h" is the closest concurrence with "H" compared with something else in the model. "SsPEEhH" incoming signal is compared with all the models kept in the user's pattern. The result will be presented in the form of a model, for which the minimal discrepancy between an incoming signal and a model is found. The global mark for differences in the route is just a sum of local distances between shotof a signal and that of a model.



Picture 1 –An example of the DTD's route

The following restrictions are imposed on the process of comparison:

The route of analysis cannot go back in time [2,3].Every shot of an incoming signal and that of a modelshould be used during the comparison.

Local marksfor the concurrence are united by adding them to the current global divergence.
Letusprovideforthetypesofanalgorithm.

We will show examples of an algorithm in a pseudo code presenting C-like language.

Symmetric algorithm

Formula to calculate the minimal global mark:

$$D(i,j) = \min \begin{bmatrix} D(i-1,j-1) \\ D(i-1,j) \\ D(i,j-1) \end{bmatrix} + d(i,j)$$

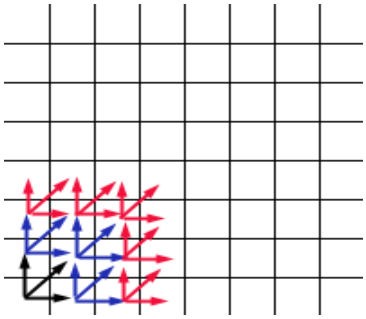
where$D(i,j)$is the global mark for the(i,j) point, and$d(i,j)$is the local one.

Algorithm of searching for a minimal global route:

1. To calculate the zero column starting from the bottom cell. The global mark for this column is equal to that of local one. In that casethe global mark for every following cell is equal to the local mark for this column plus the global mark till the column which is under it.
2. To calculate the global mark for the first cell of the next column, having added the local mark with the global mark for the lowest cell of the previous column.
3. To calculate the global mark for the residuary cells of the current column. For example, for the (i,j)point the local mark for the (i,j)point plus the minimalglobal mark from $(i-1,j)$, $(i,j-1)$or$(i-1,j-1)$.

4. The current column becomes the previous one and everything is repeated starting from the second step until all the columns are calculated.
5. The global mark is the value saved in the very top cell of the last column.

Graphical representation of an algorithm:



Picture 2-Symmetricalgorithm

Pseudo code of the described process is presented below:

```
calculate first column (predCol)
forі=1 to number of input feature vector frames
{
curCol[0] = local cost at (i,0) + global cost at (i-1,0);
forj=1 to number of template feature vector frames
{
curCol[j] = local cost at (i,j) + minimum of global cost
at (i-1,j), (i,j-1) or (i-1,j-1);
}
predCol = curCol;
}
minimum global cost is value in curCol[number of template feature
    vector frames]
```

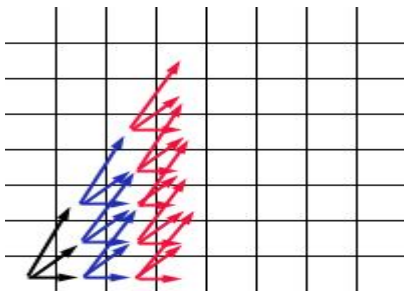
Asymmetric algorithm

Formula to calculate the minimal global mark:

$$D(i,j) = \min \begin{bmatrix} D(i-1,j-1) + 2d(i,j) \\ D(i-1,j) + d(i,j) + d_h \\ D(i,j-1) + d(i,j) + d_v \end{bmatrix}$$

Where $D(i,j)$ is the global mark for the (i,j) point, and $d(i,j)$ is the local one; suitable values for d_v и d_h can be found by experiment.

Graphical representation of an algorithm:



Picture 3-Asymmetric algorithm

Algorithm of searching for the minimal global mark has become more complicated than it was in the symmetric version. Hence, it will be much easier to explain the process in the pseudo code than to describe it with the help of words.

```
predCol[0] = local cost at (0,0);
for i=1 to number of input feature vector frames
{
curCol[0] = local cost at (i,1) + global cost at (i-1,0);
for j=1 to (minimum of number of template feature vector
frames and 2i+1)
{
store j in hoghestJ;
if the cell (i,j) is a special case
pair
{
{
if it is row 1
{
curCol[j] = local cost at (i,j) + global cost at
(i-1,j-1);
}
else
{
        if the cell (i,j) is lower of the special case
curCol[j] = local cost at (i,j) + minimum of
global cost at (i-1,j-1) or
(i-1,j-2);
}
else
{
curCol[j] = local cost at (i,j) + global cost
at (i-1,j-2);
}
}
}
else
{
if it is row 1
{
curCol[j] = local cost at (i,j) + minimum of
global cost at (i-1,j) or (i-1,j-1);
}
else
{
curCol[j] = local cost at (i,j) +
minimum of global cost at (i-1,j),
(i-1,j-1) or (i-1,j-2);
}
}
}
predCol = curCol;
}
        minimum global cost is value in curCol[highestJ];
```

Conclusion

To analyze speech it is necessary to transform it into the form clear for the computing system. This may be a digital form, spectral representation, representation with the help of analog electric signals, etc. Because of the fact that in the given paper only the process of modeling systems of speech analysis on a personal computer is described, merely one type of sound's representation is considered. To represent an acoustic signal in a digital form discrete amplitude representation is used practically in all the systems dealing with sounds. As it is known sound presents longitudinal compression-expansion waves propagating in acoustically conductive environment. By means of sound recording devices (microphone) it is transformed into an electrical signal, variations of which repeat sound vibrations.

References

1 Grobman M. Z., Tumarkin V. I., Revealing hidden periodicities and the formant analysis of speech. Images recognition: theory and applications. Science, Moscow, 1977.

2 GoldensteinSiome, Time Warping of Audio Signals, University of Pensilvania, VAST Lab. http://www.graphics.cis.upenn.edu/

3 Wrigley Stuart, Dynamic Time Warping, Internet, http://www.dcs.shef.ac.uk/~stu/com326