A n t t i   I i v o n e n

University of Oulu
Institute of Phonetics

AUTOMATIC RECOGNITION OF SPEECH SOUNDS BY A DIGITAL COMPUTER

Three contributions concerning the discrimination of the momentan spectrums of some selected Finnish and German sounds

The main difficulties in the speech recognition may be listed in the following way:

1. Which should be the basic linguistic units to be recognized: sounds (allophones), phonemes, segment combinations, syllables, words?
2. Should the output text be written ortographycally? How then the problem of the differences between the phonemic form of an utterance and the ortography should be resolved?
3. If the word is chosen as basic units for the recognition, how one should resolve the problem of the grammatical flexion (e.g. in Finnish)?
4. How can the recognition automation decide, where there is a boundary between two words or two sentences?
5. How can the automation decide that e.g. the pause during a long voiceless stop consonant is not a boundary?
6. How can the automation discriminate the tonal and croneme classes in laguages, in which they are linguistically relevant?
7. The automation should not take into account the irrelevant noise; one must regard also the noise produced by the automation itself.

8. How to localize the points in the speech continuum, which the recognition can be based on; is there one special acoustic segment (or a momentan spectrum) for every sound, which is characteristic for the sound?

9. It has been shown that segments, which are l i n g u - i s t i c a l l y i d e n t i c a l, can be acoustically different. The differences are due to following factors: (1) The same speaker can not produce two exact similar sounds, because the conception of the identity is a human abstraction. (2) Different speakers produce linguistically the same sound in a different way. (3) Linguistically the same sound can be modified acoustically by the word prominence, sentence prominence, environment, emotional factors, speech tempo, dialectal background of the speaker, speech defects, huskiness, and so on.

1o. L i n g u i s t i c a l l y d i f f e r e n t sounds can be acoustically similar.

11. Should the phonotactic structures (Sigurd) or the characteristic sequencies (Pike) of a language be regarded when creating the recognition program?

12. The technical problems form one great part of the speech recognition. They concern the m e c h a n i c a l s o l u t i o n s and the r e c o g n i t i o n p r o - g r a m.

1. Vowel recognition based on some selected vowel variables and discriminant analysis.

The probability of correct identification of the acoustically close German vowel phonemes /i:, I, e:, ɛ, y:, and Y/ on the basis of spectrographic input data and the discriminant analysis (literat. 1, 2, and 3) was calculated. One male speaker were used. Following variables were measured: the frequencies of the four first formants (F1... F4), their amplitudes (L1...L4), the amplitude of the zero (minimum) point between F1 and F2 (here called LZ1) and that between F2 and F3 (LZ2), and the duration of the vowels.

The probability of correct identification was 94 per cent on average. The highest identification probability was shown by the phoneme /e:/ (98,9 %) and the lowest by the phoneme /Y/ (85,7 %). The sounds were picked up from sentences read by the informant.

In the real classification procedure which was connected to the probabilistic recognition program 6 identifications were false out of 103 possible. The order of the significance of the variables studied regarding their discriminatory power was F2, LZ1, F1, F4, duration, L1, F3, L4, LZ2, L3. - One must take into account the possibility that two variables, the discriminatory power of which is good, will correlate with each other. In this case the better one is placed in a high position in the list, but the other one comes later than its real discriminatory power implies, because the correlation is taken into account. If the better variable was not considered, the weaker variable would perhaps take its place (if the correlation is strong enough). This may explain the fact that F3 comes after F4 (the correlation of F2 with F3 is strong concerning the vowels studied).

The energy minimum between F1 and F2 (LZ1) had a good discriminatory power. This showes that in the acoustic signal there can be cues, which are available in the automatic recognition, such cues, which need not to be relevant for perception (cf. Tillmann, p. 149).

2. Recognition based on the discrimination of the numerical models of sounds.

In the second experiment the input data of the recognition program consisted of the numerical describers of the sounds. They were formed by using constant points in the measurement of the spectrums of sounds. Thus the describer of a sound consisted of a serie of numbers, which indicated the amplitude at constant selected frequencies. The narrow filter (with 45 Hz bandwidth) was used when producing the sections, which formed the material measured. 32 measurement points inside the range of 4 kHz were used.

The describers for 330 Finnish sound manifestations were calculated. These sounds were representatives for 8 short Finnish vowel or 3 nasal phonemes /a, e, i, o, u, y, ä, ö, m, n, n/. 30 representatives of every phoneme type were picked up from sentences read by a single male speaker.

The data thus obtained were stored and submitted to the discriminating analysis. The measurement points were handled as variables.

The probability of correct recognition was about 60...70 % on average. One must regard, however, that the localization of the sections was (under circumstances) not very exact and the technical equipment was unfortunately not the best one.

3. Recognition based on the numerical models of sounds and a special recognition program.

In the third recognition experiment the Finnish nasal sounds belonging to the phonemes /n/ or /m/ were tried to be classified automatically on basis of the numerical describers, which are discussed in the preferring chapter.

Firstly the frequency area of 4 kHz was studied by means of 33 constant measurement points with distances of 121 Hz. The 'general' describers for /n/ and /m/ were calculated by means of the PROGRAM I (below).

The basic material consisted of 87 wide band sections (made with Kay Electric Co. Sound Sona-Graph model 6061-B). The sections were made from the target point of F2 of the nasals in single words (all possible environments were considered). The describers of /n/ and /m/ are presented graphically in fig. 1. The influence of the environment on the dental nasals (n) seems not to be very great (fig. 2).

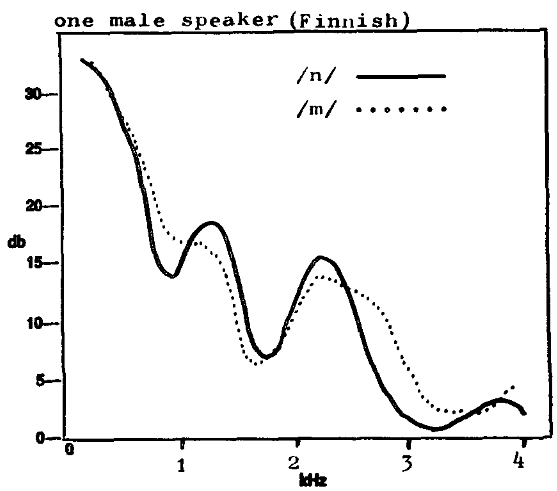

Fig. 1

Models of /n/ and /m/ phonemes. 50 + 37 wide band sections were used

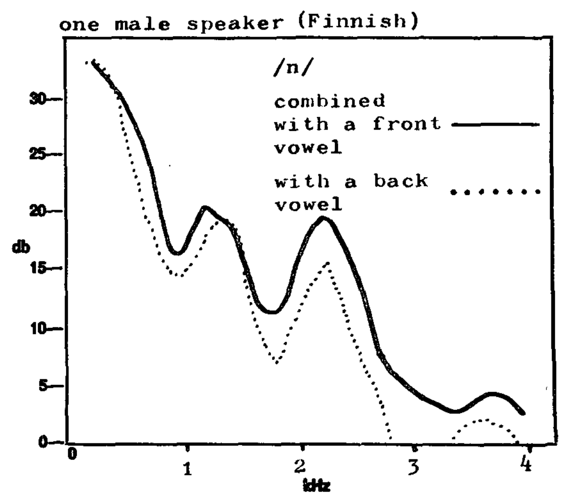

Fig. 2

Models of /n/ in different environments. Wide band sections were used.

Secondly the numerical describers were restricted so that only nine constant measurement points were considered. The nine points with the best discriminatory power were sought by means of the PROGRAM II (below).

Thirdly the 'general' numerical models for the both phonemes were calculated on basis of the nine points mentioned. The logic of the procedure is described shortly at the beginning of the program (PROGRAM III).

With the same method the numerical model of a new nasal sound was calculated (PROGRAM III), and the nasal sound was classified by compairing its model with the mean of the models of /n/ and /m/.

The main idea of classification is that the amplitudes at the nine measurement points are set on order of magnitude, and then their relative places on the frequency axis are indicated by means of the ordinal numbers (nine possibilities). The ordinal numbers are then placed one after another, so that they form one single number. This number is handled as the numerical model of a group of nasal sounds or a single nasal sound.

The classification time of a sound by means of method described here is only a fraction of that when using the discrimination analysis.

Final comments

Every language needs its own recognition program consisting of subprograms, which can be very different. That the recognition program can be worked out implies that there is a sufficient amount of acoustic knowledge about the language in question.

It is possible that the complete speech recognition doesn't succeed with the computers available, so that we must waite so long that the biological computers are at our disposal. (contin. after the programs)

PROGRAM I (programming language FORTRAN II)

```
C     COMPUTATION OF THE GENERAL MODELS FOR N GROUPS OF
C     SOUNDS: CALCULATE THE MEAN SETS FOR THE GROUPS.
C     MATERIAL CONSISTS OF MEASUREMENT VALUES AT 33
C     CONSTANT MEASUREMENT POINTS ON THE FREQUENCY
C     AXIS OF EVERY SOUND.
C     UNIVERSITY OF OULU,FINLAND
C     INSTITUTE OF PHONETICS
C
      DIMENSION IAMPLI(33),NUMBER(33),ISUM(33)
      DIMENSION AMEAN(33)
      WRITE(3,222)
  222 FORMAT('1',' ')
      IGROUP=0
  401 DO 300 I=1,33
      ISUM(I)=0
  300 NUMBER(I)=0
    1 READ(1,10)(IAMPLI(I),I=1,33)
   10 FORMAT(33I2)
      DO 200 I=1,33
      IF(IAMPLI(I)-36.00000)3,4,5
    3 NUMBER(I)=NUMBER(I)+1
      ISUM(I)=ISUM(I)+IAMPLI(I)
  200 CONTINUE
      GO TO 1
    5 DO 100 I=1,33
      AMEAN(I)=ISUM(I)/NUMBER(I)
  100 CONTINUE
      IGROUP=IGROUP+1
      WRITE(3,333)IGROUP
  333 FORMAT('0','GROUP',T8,I4)
      WRITE(3,11)(AMEAN(I),I=1,17)
   11 FORMAT(' ','MEANS',T10,17F5.1)
      WRITE(3,12)(AMEAN(I),I=18,33)
   12 FORMAT(' ',T10,16F5.1)
      GO TO 401
      END
```

The last card in a group of sounds: 999999999999...99
The last card in the program: 3636363636...36
The greatest possible value of variables (IAMPLI): 35

PROGRAM II

```
C       SEEK THE NINE BEST DISCRIMINATING POINTS ON THE
C       FREQUENCY AXIS OF THE N AND M SOUNDS. USE THE
C       NUMERICAL DESCRIBERS OF N AND M FORMED BY MEANS OF
C       THE PROGRAM I.
C
        DIMENSION AMEANN(33),AMEANM(33),ASQUAR(33),DIFF(33)
        DIMENSION BSQUAR(33),NUM(33)
C
C       CALCULATE THE DIFFERENCES OF THE DESCRIBERS OF N
C       AND M.
        :
        :         It is assumed that the describers of /n/
        :         and /m/ are stored before; they are called
        :         AMEANN and AMEANM.
        :
        DO 60 I=1,33
        DIFF(I)=AMEANN(I)-AMEANM(I)
     60 CONTINUE
        DO 61 I=1,33
        ASQUAR(I)=DIFF(I)**2
     61 CONTINUE
C
C       SET THE AMPLITUDE DIFFERENCES IN ORDER OF MAGNITUDE
        DO 421 M=1,33
        BSQUAR(M)=ASQUAR(M)
    421 CONTINUE
    423 DO 424 I=1,32
        I1=I+1
        DO 424 N=I1,33
        IF(ASQUAR(I)-ASQUAR(N)425,424,424
    425 AUX=ASQUAR(N)
        ASQUAR(N)=ASQUAR(I)
        ASQUAR(I)=AUX
    424 CONTINUE
C
C       INDICATE THE ORDINAL NUMBERS OF THE POINTS MEASURED
C       IN ORDER OF DISCRIMINATING POWER
        DO 450 I=1,33
        IORDER=0
        DO 2 M=1,33
        IORDER=IORDER+1
        IF(ASQUAR(I)-BSQUAR(M))7,7,2
      7 NUM(I)=IORDER
        BSQUAR(M)=-9999999.0
        GO TO 450
      2 CONTINUE
    450 CONTINUE
        WRITE(3,14)(NUM(L),L=1,33)
     14 FORMAT('0','ORDINAL NUMBERS',T20,33I3)
        CONTINUE
        END
```

PROGRAM III

```
C     AUTOMATIC DISCRIMINATION OF N AND M
C     UNIVERSITY OF OULU FINLAND
C     INSTITUTE OF PHONETICS
C
C     LOGIC OF THE PROGRAM:
C     1:CALCULATE THE MEANS OF THE AMPLITUDES AT THE NINE
C     MEASUREMENT POINTS,WHICH ARE THE MOST DISCRIMINATING
C     POINTS ON THE FREQUENCY AXIS FOR N AND M!
C     2:SET THE AMPLITUDES IN ORDER OF MAGNITUDE!
C     3:INDICATE THE ORDINAL NUMBERS OF THE AMPLITUDES!
C     4:FORM THE GENERAL NUMERICAL MODEL FOR N AND M
C     ON BASIS OF THE ORDINAL NUMBERS!
C     5:CALCULATE THE MODELS OF NEW NASAL SOUNDS WITH
C     THE SAME METHOD!
C     RESOLVE THE PROBLEM:IS THE NEW NASAL SOUND A N OR
C     A M? COMPAIR ITS MODEL WITH THAT OF THE GENERAL
C     MODELS OF N AND M!
C
      DIMENSION ASUM(9),AMEAN(9),BSUM(9),BMEAN(9),NUM(9)
      DIMENSION AMPLIT(9),NUMBER(9),INUMBR(9)
C
C     COMPUTATION OF THE MEANS IN THE BASIC MATERIAL
C     CONSISTING OF A SET OF N AND M SOUNDS
      WRITE(3,222)
  222 FORMAT('1',' ')
      K=1.00000
      GO TO 401
  400 K=K+1
      :
      :        The principle of calculating the means
      :        is presented in the PROGRAM I .
      :
C     SET THE AMPLITUDES IN ORDER OF MAGNITUDE
      INDIV=0
      GO TO 59
  770 K=-1.00000
      :
      :        The principle of calculating the order of
      :        magnitude is presented in the PROGRAM II.
      :
C     FORM THE ORDINAL NUMBERS:FOR EXAMPLE:THE GREATEST
C     AMPLITUDE WAS THE NINETH IN ORDER
      DO 450 I=1,9
      IORDER=0
      DO 2 M=1,9
      IORDER=IORDER+1
      IF(AMEAN(I)-BMEAN(M))7,7,2
    7 NUM(I)=IORDER
      BMEAN(M)=-9999999.0
      GO TO 450
```

```
    2 CONTINUE
  450 CONTINUE
      WRITE(3,14)(NUM(L),L=1,9)
   14 FORMAT('0','ORDINAL NUMBERS',T20,9I6)
C
C     FORM THE NUMERICAL MODEL
      MODEL=0
      MULTPL=100000000
      DO 30 M=1,9
      IPROD=NUM(M)*MULTPL
      MODEL=MODEL+IPROD
      MULTPL=MULTPL/10
   30 CONTINUE
      IF(K-1.00000)49,51,52
   51 WRITE(3,31)MODEL
   31 FORMAT('0','MODEL OF N',T15,I10)
      IN=MODEL
C
C     THE SAME PROCEDURE CONCERNING M
      GO TO 400
   52 WRITE(3,66)MODEL
   66 FORMAT('0','MODEL OF M',T15,I10)
      IM=MODEL
      MEAN=(IN+IM)/2
      WRITE(3,111)MEAN
  111 FORMAT('0','THE MEAN OF N AND M',T25,I10)
C
C     FORM THE FORM A NEW NASAL SOUND
  550 DO 330 M=1,9
  330 AMEAN(M)=0.0
      DO 334 M=1,9
  334 BMEAN(M)=0.0
      READ(1,9)(AMEAN(M),M=1,9)
    9 FORMAT(9F4.0)
      IF(AMEAN(I)-36.00000)660,888,888
  660 GO TO 770
   49 NAS=MODEL
      WRITE(3,77)MODEL
   77 FORMAT('0','MODEL OF NASAL',T18,I10)
C
C     CLASSIFICATION OF THE NEW NASAL SOUND
      INDIV=INDIV+1
      WRITE(3,98)INDIV
   98 FORMAT('0','INDIVIDUAL',T15,I3)
      IF(NAS-MEAN)801,802,803
  801 WRITE(3,900)
  900 FORMAT(' ','= M)
      GO TO 123
  802 WRITE(3,901)
  901 FORMAT(' ','= M OR N')
```

```
    GO TO 123
803 WRITE(3,9o2)
902 FORMAT(' ','= N')
123 CONTINUE
    GO TO 550
888 CONTINUE
    END
```

If the recognition of the natural languages isn't possible, we should consider the possibility of an artifical language, which would be easy to be recognized by a machine.

If the social need of the recognition automations becomes very great, it is possible that the conservative orthography of many language will disappear, and the phonematic orthography will become common.

The discriminant analysis used in this contribution has been programmed by Mr. S. Sarna in the Computation Centre of the University of Helsinki (cf.2).

Literature comments

1. Mustonen, Seppo: Multiple Discriminant Analysis in Linguistic Problems. Nordsam 64, Det Femte Nordiska Symposiet över Användning av Matematik Maskiner. Stockholm 18.-22.8.1964.
2. Sarna, Seppo: Erotteluanalyysin periaatteet ja käyttömahdollisuudet. Mimeographed copy. Computation Centre of the University of Helsinki (1968).
3. Cooley, W.W. and Lohnes: Multivariate Procedures for Behavioral Sciences. New York, John Wiley and Sons (1962).
4. Tillman, H.G.: Akustische Phonetik und linguistische Akustik. Phonetica 16: 143-155 (1967).
5. Tillmann,H.G., Heike,G., Schnelle,H. und Ungeheuer,G.: Dawid I - ein Beitrag zur automatischen "Spracherkennung". 5e congres international d'acoustique. Liège 7-14 septembre 1965.