International Council for the
Exploration of the Sea

ICES 1991/ C: 22
Hydrography Committee
Ref: D

VOICE CONTROLLED FORM FILLING

By
Trygve Gytre
Inst. of Marine Research
P.O. Box 1870 / 72
N-5024 Nordnes

ABSTRACT

During field work in the sea - in particular when collecting data like weight, length, sex, stadium , etc. from fish samples, the observations are first manually noted in a standardized sampling form. One popular form used for fish studies at the Inst. of Marine Research is the "Individual (V)" - form. Since the observer needs both his hands to handle the specimens, the V form is normally filled out by an assistant.

In due time another assistant will transform the infomation contained in the form to an ASCII file by "punching it" into a computer. After this the data are available for general use.

Lots of time and personel could be saved if an observer can generate error free ASCII files from his data immediately and alone. His hands are occupied and hardly clean enough to touch a computer keyboard, but his voice is clean and vacant.

In an introductory project sponsored by the Norwegian Fishery Research Council a word recognition system for form filling has been designed. The system involves voice interfacing components, a PC with "speech board" and an application programme that accepts voice input and also returns spoken words for user feedback.. The system has been used to generate ASCII formatted "V" - forms by dictation.

The system shows good ability to recognize multi-syllable words. Short words are not so easily identified. Identification of words from a vocabulary with a mixture of short and long words show a first try failure rate of 5-10 %.

By using a vocabulary specially adapted for voice regognition, the failure frequency can be reduced .

THE "V"- FORM

Fig. 1 shows an authentic "V"- form which has been manually filled by an observer at the Inst. of Marine Research.

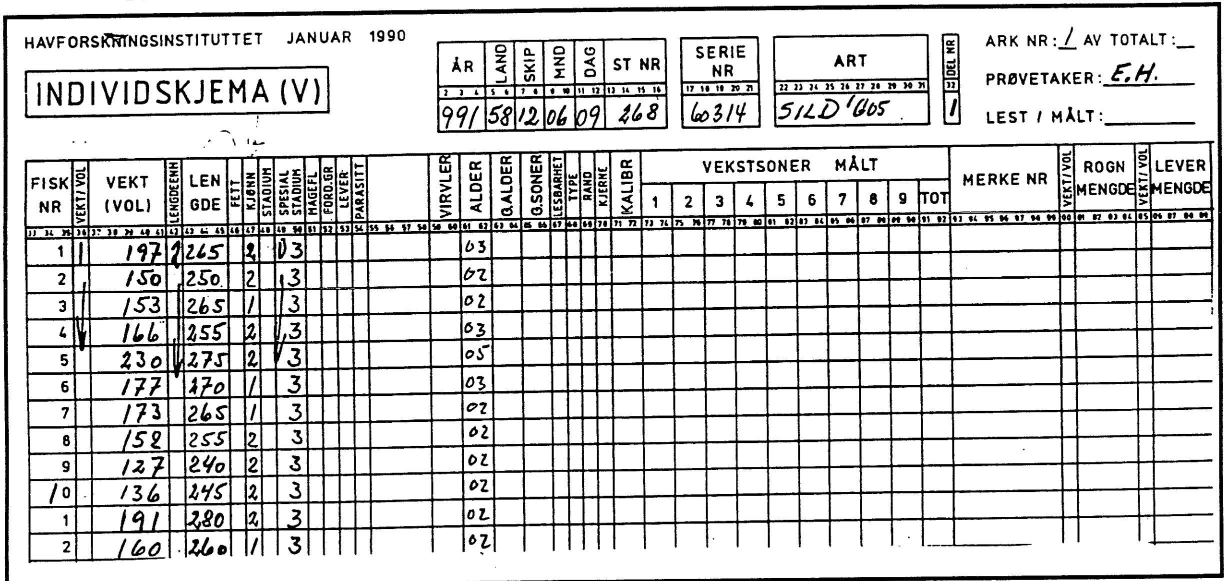
The form, which uses the Norwegian language, has a fixed heading for year (år), country (land, ship (skip) month (mnd),day (dag), station number (st nr), series no. and species (art).

During the fish sampling process the observer fills in weight or volume, length, fat level, sex, stadium, stomach filling, liver,parasite, number of whirls, age...... etc.

A complete filling of this form typically takes three days:

Day one : Capture fish , record length, weight, sex, state. Filet the fish.
Day two : Count and record number of whirls, remove otholites.
Day three: Read otholites, record age.

In most applications only parts of the form is used.



HAVFORS~~KN~~INGSINSTITUTTET JANUAR 1990

INDIVIDSKJEMA (V)

ÅR	LAND	SKIP	MND	DAG	ST NR	SERIE NR	ART	DEL NR
991	58	12	06	09	268	60314	SILD'G05	1

ARK NR: 1 AV TOTALT:__

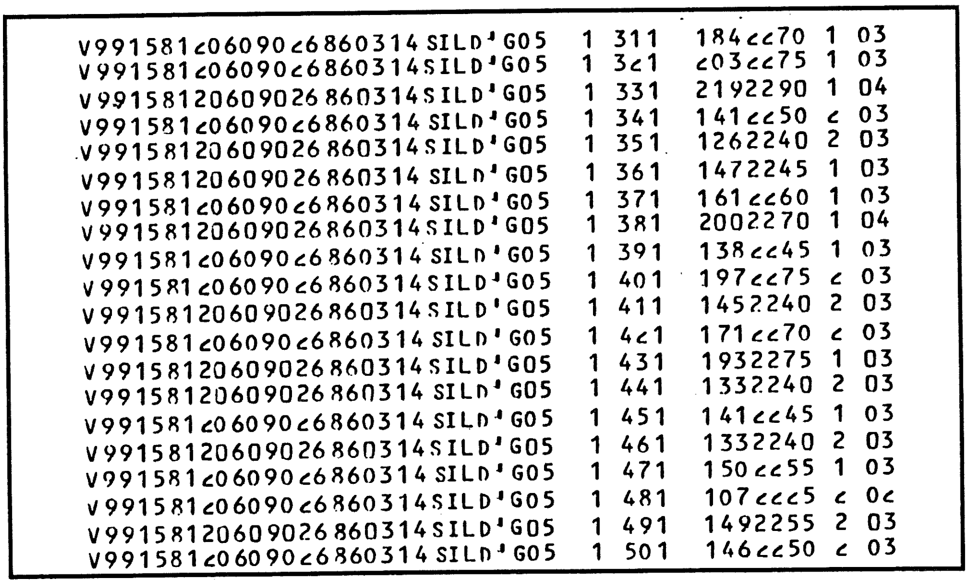
PRØVETAKER: E.H.

LEST / MÅLT:________

FISK NR	VEKT/VOL	VEKT (VOL)	LENGDEENH	LEN GDE	FETT	KJØNN	STADIUM	SPESIAL STADIUM	MAGEFL	FORD.GR	LEVER	PARASITT		VIRVLER	ALDER	G.ALDER	G.SONER	LESBARHET	TYPE	RAND	KJERNE	KALIBR	VEKSTSONER MÅLT										MERKE NR	VEKT/VOL	ROGN MENGDE	VEKT/VOL	LEVER MENGDE
																							1	2	3	4	5	6	7	8	9	TOT					
1	1	197	2	265		2		3							03																						
2		150		250		2		3							02																						
3		153		265		1		3							02																						
4		166		255		2		3							03																						
5		230		275		2		3							05																						
6		177		270		1		3							03																						
7		173		265		1		3							02																						
8		152		255		2		3							02																						
9		127		240		2		3							02																						
10		136		245		2		3							02																						
1		191		280		2		3							02																						
2		160		260		1		3							02																						

Fig. 1 Section of a manually filled "V"- form used by the Inst. of Marine Research

After the data have been punched, the computer will transform the data to ASCII files . Fig. 2 shows parts of a printout for a typical ASCII file containing V- data from North Sea herring.



```
V991581206090268860314 SILD'G05   1 311   1842270 1 03
V991581206090268860314SILD'G05    1 321   2032275 1 03
V991581206090268860314SILD'G05    1 331   2192290 1 04
V991581206090268860314 SILD'G05   1 341   1412250 2 03
V991581206090268860314SILD'G05    1 351   1262240 2 03
V991581206090268860314 SILD'G05   1 361   1472245 1 03
V991581206090268860314 SILD'G05   1 371   1612260 1 03
V991581206090268860314SILD'G05    1 381   2002270 1 04
V991581206090268860314 SILD'G05   1 391   1382245 1 03
V991581206090268860314SILD'G05    1 401   1972275 2 03
V991581206090268860314SILD'G05    1 411   1452240 2 03
V991581206090268860314 SILD'G05   1 421   1712270 2 03
V991581206090268860314SILD'G05    1 431   1932275 1 03
V991581206090268860314 SILD'G05   1 441   1332240 2 03
V991581206090268860314 SILD'G05   1 451   1412245 1 03
V991581206090268860314SILD'G05    1 461   1332240 2 03
V991581206090268860314 SILD'G05   1 471   1502255 1 03
V991581206090268860314SILD'G05    1 481   1072225 2 02
V991581206090268860314SILD'G05    1 491   1492255 2 03
V991581206090268860314 SILD'G05   1 501   1462250 2 03
```

Fig. 2 Section of a typical V- form after it has been punched and converted to an ASCII file

BASIC VOICE RECOGNITION THEORY

Physically the human speech is a complex, rapidly changing pressure waveform.

Major changes between different acoustic states in that waveform usually correspond to transitions between the phonemes (major speech sounds) in the spoken words. Acoustically the same phonemes vary widely when spoken by different people.

Therefore most voice recognition systems based on recognition of individual phonemes must be "trained" with the actual voices before they can be brought to work.

An artificial voice recognition process starts by detecting the unknown voice sounds by a microphone and feeding the output signal to a processing system. Being initially analogue, the speech waveform must be digitized before computer processing can be carried out.

It is normally assumed that the speech waveform remains relatively unchanged over10 ms. periods or "frames" Most speech recognition system will first divide the speech waveforms into frames and digitize the signals in each frame 60-200 times. This will generate a stream of digital, computer understandable data. After being entered to the computer, the information in each frame will be processed according to a strategy.

PROCESSING STRATEGY

The two most used strategies today are time-domain processing and linear predictive coding. The time domaine processing simulates the processes that take place in coclea in the human inner ear. Fig. 3 illustrates the strategy. Using either analogue or digital technique the energy in each frame is spectral analyzed and sorted into 8 - 32 energy bands . The relative energy distribution in the selected frequency bands define a sound "model" for that particular frame. The models are stored in the memory for comparision with similar frame templates obtained during previous training.

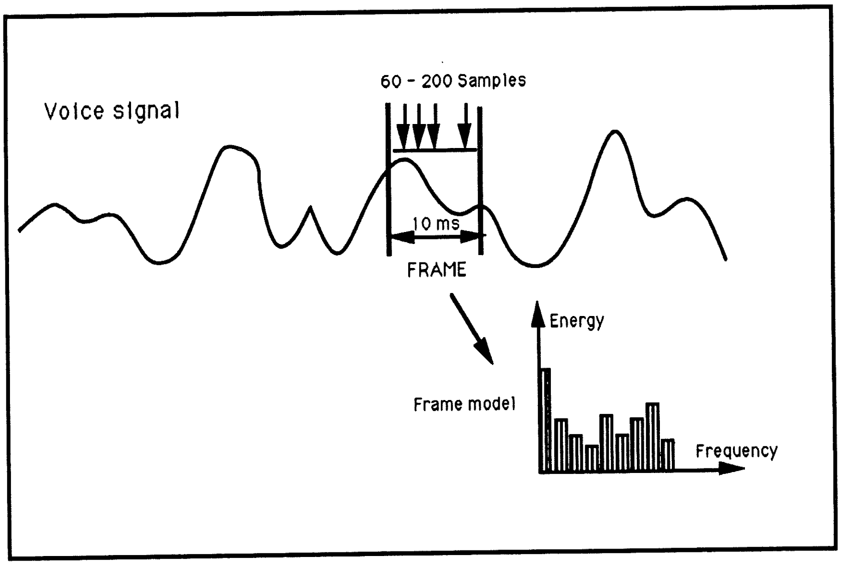


Fig. 3 Time domaine voice processing:
The analogue speech waveform is divided into appr. 10 ms wide "frames" Each frame is digitized and its energy specter is computed. The relative energy distribution becomes basis for frame models .

In linear predictive coding, the processing will model the human vocal tract as a linear acoustic filter that is excited with random signals from a generator. The responses of the filter are determined by a set of time varying coefficients. Signals from the generator and from the speaker`s microphone are compared in a processor. A set of coeffisients that make the output from the filter to match with the microphone signals are determined from the least-squares fits.The coeffisients, which represent models, are compared with similar coefficients obtained during training.

Fig . 4 illustrates linear predictive coding.

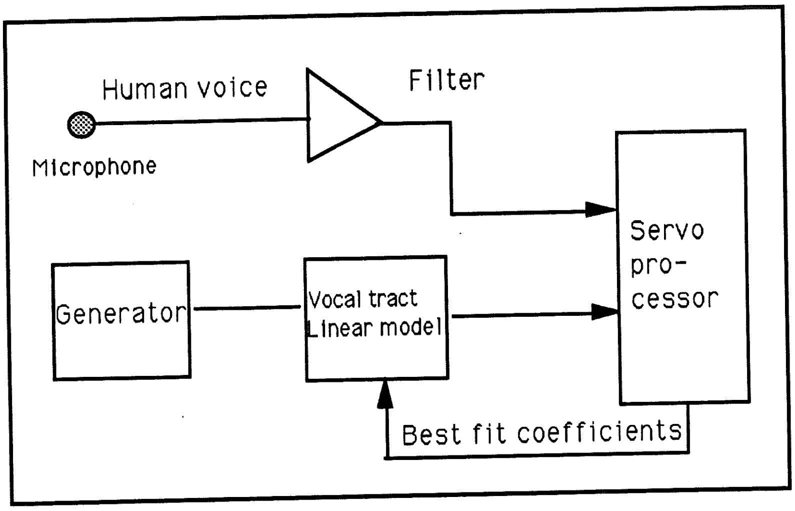


Fig. 4
Linear predictive voice recognition.
A processor compares human voice with artificial voice from a model of the human vocal tract.
Computed coefficients needed to make artificial signal fragments equal to those received by the microphone represent sound models.

SYSTEM DESIGN

Fig. 5 shows a block diagram for the complete system.

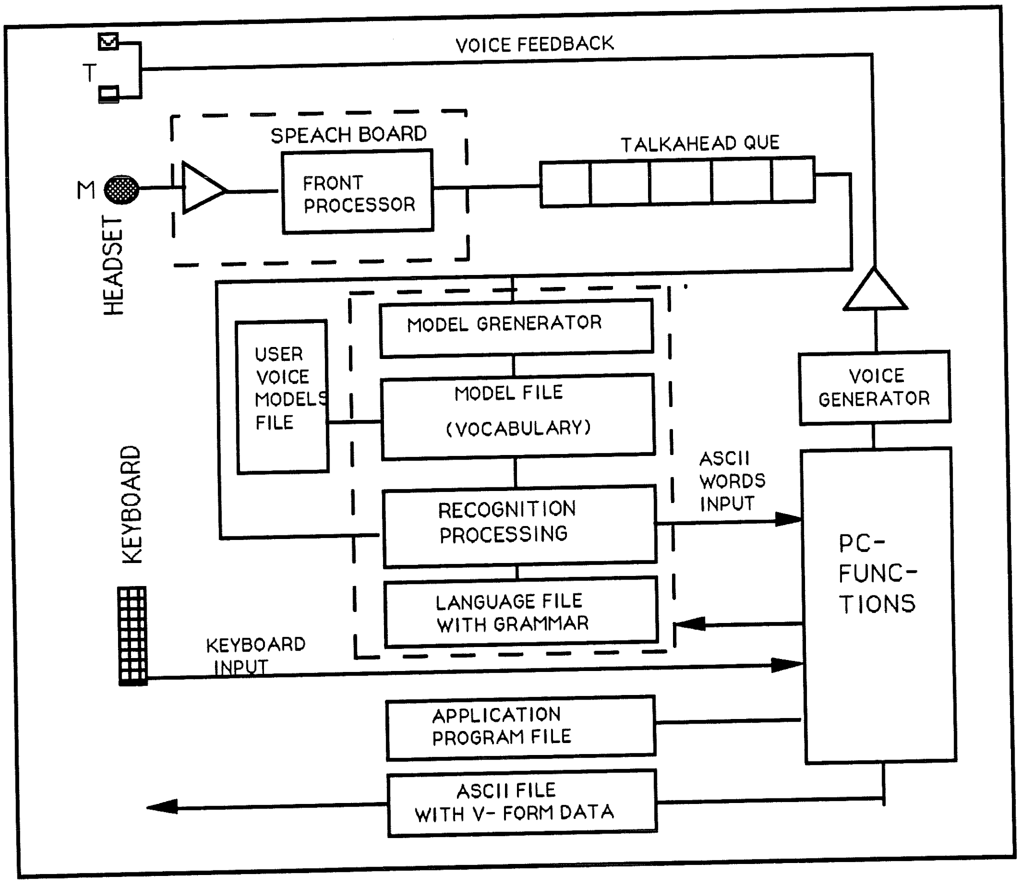


Fig. 5 Block diagram for the complete voice recognition system used to generate ASCII files from dictated information

The dominant part of the voice controlled V- form filling system is a standard AT type PC which has been expanded with a plug-in speech board (Voice Scribe 1000) made by Cherry Inc. USA. Data and commands to the system can be input both from microphone and from keyboard. The system can output feedback and information via audible spoken words from a word generator, from the PC- screen and from the standard P.C. output ports.
In order to talk to the system and to perceive the spoken feedback from the word generator in a convenient way, the observer carries a headset with microphone and two earphones.

Before a operator can use the system he must "train" it with his own voice by dictating each word used by the V- form or needed for commands 3 - 4 times. In this contex a word is defined as a continuous utterance spaced by at least 0.1 s of silence in both ends.
The system generates and records a model for each individual word. A separate file is kept for each registered user. To to fill the V- form, a vocabulary of appr. 25 words is needed.
The words are numbers 0-9 , the words used in the form and the command words "ENTER" , "HOPP" (Jump) and "END"
Words detected by the microphone are digitized and analyzed in the speech board. The analyzed data are compared with models of the valid vocabulary words. When a word is accepted to be "equal enough," the system generates an ASCII code to the PC. The code activates the application program.
Via the voice generator and earphones the computer tells the observer what information it wants for next input. It also repeats inputs for control and possible correction.
The end product is an ASCII file with identical structure to the one shown in fig. 2.
Fig. 6 illustrates a typical dialogue between user and system during the V-form filling.

PC -VOICE (Primary)	OPERATOR VOICE	PC -VOICE (Feedback)
FISH NUMBER	ONE ENTER	ONE
WEIGHTVOLUME	ONE ENTER	ONE
WEIGHT	ONE NINE SEVEN ENTER	ONE NINE SEVEN
LENGHT UNIT	TWO ENTER	TWO
LENGHT	TWO SIX FIVE ENTER	TWO SIX FIVE
SPECIAL STADIUM	THREE ENTER	THREE
AGE	ZERO THREE ENTER	ZERO THREE
FISH NUMBER	JUMP FISH NO ONE ONE ENTER	ONE ONE
	LENGHT TWO FOUR ZERO ENTER	TWO FOUR ZERO
	END	

Fig. 6 Dialogue during form filling.

A Normal vocal input

B Error corrction

The standard station data headings are most easily filled in via keyboard before or after entering the biological information.
The system starts by asking for "fish number". (The actual words used by the programme are Norwegian and the dialogue has been translated)
The observer answers "one" and adds "enter". When the computer hears "enter", it repeats the perceived word for user`s control. Then the computer steps to next column and says "weightvolume". The observer continues by saying "one enter" and gets a "one" in return etc.
In this way the observer can enter all data for all his fishes or, if more conveniently, only parts of the data at a time.
When one line is finished, the system automatically jumps to next line.
To stop and close the file, the observer says "END".

Sooner or later erroneous data will occur. The observer can correct faulty data entries by jumping to the position with erroneous data and overwrite the positions with correct data.
The lower part of fig. 6 shows how a correction can be made. The observer has discovered an error in length for fish number 11. He jumps to the fish number column by the command "JUMP FISH NUMBER" The computer says "FISH NUMBER" The users selects line 11 by saying "one one enter". He receives the feedback "ONE ONE" Then the user says "LENGHT TWO FOUR ZERO ENTER" The computer responds with "TWO FOUR ZERO" and the error has been corrected.

Results:
The performance of the voice controlled system is now being tested. Tests show that when all data have been correctly perceived by the system, error free ASCII files are created.
However, erroneous inputs or system rejections are easily introduced through undistinct word pronounciation and through background noise.
Short syllable words are more difficult to detect than long words, and pulsed noise makes more harm than a constant background noise. In its present state appr. 5 - 10 % of the pronounced words are typically rejected or misunderstood at first try. Experiments with improved components and selection of special words suited for voice recognition will be continued.
Field test in natural noisy environments on board one of the Institute`s research vessels is planned in spring 1992. In the meantime the system will be further refined.

Discussion:
Word recognition is still at an introductory stage compared to the expectations, but it is a fast growing field in modern computing. When a deeper understanding of the way information is carried from person to person through the spoken word has been made, better models and more sophisticated components will probably make speech input to computers as natural as keyboard inputs.
In its present stage sound recognition can best become a practical tool for form filling if a special application oriented dictatating language can be designed. The observer could f. inst. say "oneandone",twoandtwo", threeandthree" etc. instead of "one", "two","three"......
A special vocabulary for form filling will be designed.

Ref.: Dragon Key users manual Release 2.10
The Cherry Corporation 3600 Sunset Av. Vaukegan Ill. 60087.