SPECTROGRAM READING AND EXPERT METHODS FOR ACOUSTIC-PHONETIC SPEECH SIGNAL DECODING

Nina Zinovyeva

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

This paper presents the results of a large series of experiments in reading spectrograms of Russian utterances. Our experiments have enabled us to reveal the most general principles of human speech behaviour in spectrogram processing and expert acoustic-phonetic decoding strategies. We discuss here these aspects of human expertise and also address the problem of expert knowledge implementation in designing speech recognition algorithms, e.g. the algorithms for segmentation of speech wave into segments corresponding to phonemes.

INTRODUCTION

As it was stated in a recent series of papers /I/ - /5/, the experiments in spectrogram reading demonstrated the richness of phonetic information that can be derived from the most widely used three-dimensional /frequency-time-intensity/ visual display of the signal produced by visible speech spectrographs. It was also pointed that "rules for extracting and interpreting this information can be explicitly formulated" /2/ and thus used to improve the segmentation and labeling performance of present speech recognition systems /3/, /5/. It is also worth mentioning that despite the other-than-auditory modality of speech signal processing, the spectrogram reading is of particular interest as it can provide some useful insights into the human speech perception as such /I/.

Here we report on the results of a long-term investigation in reading spectrograms of Russian utterances. The experiments have been conducted at the Philological Department of the Moscow University since 1979. At the very beginning of the research the participants were not skilled spectrogram readers, so one of the goals of our study was to acquire acoustic-phonetic decoding competence in dealing with visual speech signal representation. That was the reason for rather simple experimental tasks set on an early stage of the work and their gradually increasing complexity in the following experiments. It was achieved by using more complex speech units /from isolated words of limited vocabulary and their syntactically and semantically anomalous combinations to nonsense words and nonsense phrases, syllables, extracted from words and phrases and so on/, by increasing the number of speakers /on the whole the utterances of I4 speakers were examined during our research/ and complicating the conditions of the experiments /using different "time windows" with duration ranging from 300 to 50 msec, noising speech signals etc./.

In our investigation we used to read wide-band spectrograms produced on a "Kay Sona-Graph" /Model 7029A/ with different frequency ranges. About 800 spectrograms were analyzed in total. In each experiment were taking part from 3 to 4 human readers, who in course of the research /during first 2-3 years/ mastered the skill of spectrogram acoustic-phonetic interpretation to the highest degree. We shall further refer to them as experts.

The results of acoustic-phonetic decoding achieved in our experiments were as follows: the skilled experts were able to correctly transcribe about 87% of segments with an average I,2I labels produced to each segment in the case of isolated word interpretation and about 83% with an average of I,2I labels for the connected speech.

It would be interesting to compare our results with those obtained on the basis of other languages. It was reported /I/,/3/ that for American English the mean accuracy of labeling ranges from 80% to 90% with an average of I,53 labels to each segment. For French the first measurement is approximated to 85% and the second - to I,5 /4/, /5/.

The comparison of these results makes it clear, that they are very similar in the first measurement, which reflects the accuracy of phonetic interpretation of spectrograms, and differ in the second, reflecting the ambiguity of phonetic decisions. We suppose that this difference is due to the different phonetical and phonological structures of languages under discussion, specifically to the different numbers of vowel phonemes. Vowel segments, highly influenced by the surrounding context, are more ambiguous than consonants, but relatively small set of alternative phonetic labels for vowel identification in Russian decreases the ambiguity of phonetic decisions.

The close examination of our results revealed some other factors, which influence experts' performance in spectrogram reading. This performance depends on the skill level of the expert due to the training period of spectrogram reading, on the type of analyzed speech material /connected speech versus isolated words/, on using short "time window" and on the speaker's specific pronunciation features /foreign accent or speech deficiency/. At the same time the accuracy of acoustic-phonetic decoding practically does not depend on speaker's voice quality.

It is worth mentioning, that dealing with spectrograms the experts did not make precise measurements of spectral parameters, because measurement process increased difficulties in reasoning about spectrograms and tempered the results.

All the facts mentioned above, as well as the close examination of the protocoles provided by the experts and of the tape-recorded discussions which they conducted during some spectrogram reading sessions enabled us to formulate the most general principles of speech spectrogram acoustic-phonetic interpretation.

THE GENERAL PRINCIPLES OF ACOUSTIC-PHONETIC DECODING

We have formulated four most important principles of spectrogram acoustic-phonetic decoding /APhD/. It should be pointed out that we have revealed them from our own experience, but the later analysis of the literature on the problem has shown that practically all of them are somehow mentioned in the papers of other researchers /5/ - /9/. This leads us to conclude that these principles characterize experts' speech behaviour as such, unrelated to the language structure and perhaps even to the speech perception modality.

I. The phonetic identification can't be deduced immediately from the continuous acoustic-parametric representation of the signal. There exists an intermediate level of speech signal processing which serves as a kind of bridge across the representational gap between acoustic substance and it's underlying phonetic form /8/. The acoustic information on this level is described in the most compact and abstract manner, without absolute numerical measurements of spectral parameters. Such qualitative descriptions suppose the detection from spectrogram the most important and closest to phonetic categories acoustic properties, free from the signal variability due to extralinguistical sources. We believe that the training period for spectrogram reading is mainly connected with evolving in expert's mind this specific interface device for an other-than-auditory modality.

Nowadays almost all researchers in the field assent to the idea of existence of this specific intermediate representational level in speech processing. The units of this level are called acoustic cues or descriptors /I/ - /6/. Our attempt to sketch the system of these units is represented in the section below.

2. The APhD is a highly active process combining two processing directions: bottom-up and top-down. It means that the lower speech representation level is analyzed from the point of view of higher level units. The acoustic-parametric representation is judged by the set of acoustic cues existing in expert's mind. The interpretation of these acoustic cues is the result of producing and gradual decreasing of a number of phonetic hypothesis. That is why the results of spectrogram reading depend in particular on the number of units in different phonetic classes.

3. The general procedure of APhD is divided in two separate stages: a) segmentation by partial sound specification corresponding to the manner-of-articulation categories and b) identification of the place-of-articulation features /in larger sense including the front-back and high-low qualities of vowels/. It should be pointed out that segmentation does not precede labeling but is conducted by partial recognition of sound stretch. At the same time segmentation preceeds full phonetic identification because contextual evidences are used to determine the place-of-articulation features. At this stage the segment boundary placement may be refined but in any case segmentation is resulted from recognition and includes various procedures for detecting and extracting from the sound wave groups of sounds different in manner of articulation or the so called "broad phonetic classes" /I/, /3/.

4. The APhD is highly context-dependent process. The use of contextual evidences in transformation of acoustic cues into phonetic categories is obvious and generally acknowledged. But we believe that contextual information is important as well in measurement-to-descriptor mapping. The experts are not aware of this because they reach an intermediate speech processing level by unconscious mechanisms of human visual system. But our current experiments with digital representation of the signal have demonstrated the significance of contextual information at the very first stages of APhD.

THE SYSTEM OF ACOUSTIC CUES FOR PHONETIC UNITS RECOGNITION

In the present section we attempt to sketch the inner structure of the acoustic cue level /AC-level/ and it's relations with the preceding and subsequent levels of speech signal processing. We suppose that this structure reflects the expert APhD strategy.

In examination of the spectrograms the experts proved to be highly efficient in detecting some "primitive visual objects" or PVO /7/ that serve as the basis for AC-descriptions. The acoustic cues carry information about presence/absence of PVO, their sequential relations, and about duration, intensity and frequency modifications of PVO. The term "modification" in our case means some qualitative /not quantitative/ characteristics of PVO, described as being "high/mid/low", "long/short", "strong/weak" /3/, while relative characteristics are described in terms of "higher/lower", "stronger/weaker" and "longer/shorter". The information about PVO's changes through the time, reflecting in their spectral trajectories, is also used to achieve AC-descriptions.

The AC-level in turn can be roughly divided

into three sublevels: AC-I, AC-2, AC-3. These sublevels are differentiated according to their place in the whole analysis and the relative degree of approximation to the preceding and subsequent levels /fig. I/. Thus units of AC-I are closer to the acoustic-parametric representation, while AC-3 units are closer to the phonetic level. The AC-I analysis is practically independent of the phonetic structure of a certain language, but is involved in speech/non-speech detection of the acoustic signal. On the contrary, the AC-3 analysis highly depends on the phonetic system of a language. In this respect the AC-2 level is in an intermediate position.

The aforementioned sublevels are interconnected because higher level descriptions as a rule incorporate units of the preceding level. In addition, within each sublevel AC-units differ in their physical nature /duration, intensity or frequency/.

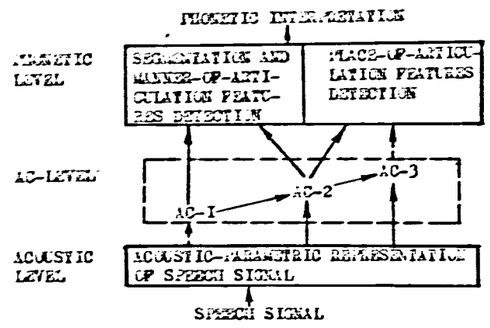

Fig. I
THE GENERAL SCHEME OF THE APhD
We have marked with broken line those components which we believe to be of paramount importance and complexity in view of speech recognition.

The presence/absence information about PVO and their combinations constitutes the very first layer of AC-I. Using these cues the speech wave is splitted into primary subphonetic segments /PSS/ such as "voiced closure", "unvoiced closure", "voiced noise", "unvoiced noise" and "vocal segment" which incorporates both vowels and sonorant consonants. AC-I also contains the sequential cues which make it possible to combine some non-vocal PSS /such as closures and following them noises/ into aggregate segments. Further on the durational AC-2 are used on these segments. In this case durational AC-2 presuppose relative estimation of lengths of PSS in terms of "longer/shorter". This procedure enables to identify stops, affricates and stop-fricative clusters with different place of articulation of their components /such as [kʃ] /.

Vocal segments, including vowels and sonorants, are estimated by intensity and frequency AC-2. These acoustic cues are rather multiple as a result of the necessity to consider different positions of sonorants within the vocal segments and different frequency locations of the formants due to the nasal/non-nasal sonorant discrimination. At the same stage the duration AC-2 are used to detect vibrants. In this case the AC-2 are formed according to the expert knowledge of minimal possible duration of the PSS.

The procedure described above results in extracting segments corresponding to the manner-of-articulation categories. All the obtained pieces of speech wave are estimated by durational AC-2, detecting segments that are likely to contain more than one phone of the same manner of articulation /e.g. [aj] /. For these segments /and for the segments corresponding to stop-fricative clusters as well/ boundary placement decisions are improved on higher levels.

The first stage of expertise concerning segmentation by identification manner-of-articulation categories doesn't comprise any significant difficulty for the experts /the accuracy of segmentation is about 98% correctly placed boundary markers/. But in order to have the computer mimic the segmentation performance of the experts, we need to evolve some special device for automatic extraction of the PVO. This problem is very difficult to solve. It lies beyond our competence /and beyond the field of phonetics/ as it is related to the mechanisms of human visual object recognition.

A more complex intellectual task for the spectrogram readers is to derive place-of-articulation values from the AC descriptions. To achieve this aim they at first use intensity and frequency AC-2, describing contextually independent modifications of consonants /mainly of their noise components/. But very often it is rather difficult and sometimes even impossible to specify accurately the place of articulation for consonants only by their intrinsic characteristics. In such cases the experts use AC-3 descriptions, carrying information about PVO's /mainly vowel formants'/ changes through the time. The AC-3 analysis is highly contextual and presupposes parallel inference of both consonant and vowel phonetic specification, based on the same acoustic evidences.

Formant trajectories are described qualitatively in terms of movement directions, relative spectral locations and frequency ranges, shape and slope. The phonetic decisions are deduced according to the expert knowledge of acoustic manifestations of the coarticulation processes in Russian which make it possible to interpret formant trajectories in terms of implied acoustic targets, closely connected with place characteristics.

To implement the rules of AC-3 inference in designing speech recognition systems it is necessary not only to reveal and formulate all the knowledge items of this level, which is not minor problem in itself, but also to describe their possible and impossible combinations, resulting in different confidence values of phonetic decisions. Besides, it is necessary to evolve quantitative methods to create qualitative descriptions, used by the experts. So, the conversation of AC-3 units into phonetic features is the second most difficult step of APhD to mimic it in the computer programs.

IMPLEMENTATION OF THE EXPERT APhD PRINCIPLES IN ALGORITHM FOR SPEECH SIGNAL SEGMENTATION

At present we devote our efforts to adapt the spectrogram reading methods to digital representation of the signal which can serve as an input into a speech recognizer. To bypass the problem of automatic extraction of the PVO, we were to select suitable digital representation for indirect interpretation in terms of AC-I. We came to the conclusion that frequency band analysis creates appropriate representation for this purpose because the information about energy balance in frequency bands can be qualified as a result of the PVO being present or absent. It proved impossible to select uniformal for all speakers frequency range division into bands, but we believe that this problem can be solved by evolving a rather simple adaptation procedure to define frequency band boundaries according to the speaker's voice quality.

In general the selection of frequency bands depends on the ranges of functioning of acoustic objects corresponding to the PVO /"voice", "FI", "FII", "FIII", "low velar noise", "mid alveolar noise" and "high dental noise"/. It is evident that in this case frequency bands would overlap because of the overlapping functional ranges of the above-mentioned acoustic objects /e.g. frequency band distribution for one of the speakers: up to 300 Hz, 200-900 Hz, 500-3500 Hz, 3500-7000 Hz/.

Using the achieved digital representation /i.e. a series of parameter vectors, reflecting frequency band energy concentrations, one every centisecond/, we have designed a segmentation algorithm, implementing all the expert APhD principles described above. The algorithm consists of different procedures /=sets of rules/ for detecting different in manner of articulation groups of phones /segmentation by recognition/. Each procedure performs goal-directed search for those pieces of speech wave that are consistent with the preconditions of its rules /active search/. The procedures include rules of interpretation of frequency band information in terms of AC-I /intermediate representation/. The algorithm doesn't perform centisecond phonetic labeling but interprets each parameter vector with respect to adjacent vectors /contextual analysis/.

We'll dwell on the last principle in more detail. The expert analysis of initial digital representation revealed that vast majority of centisecond parameter vectors were characterized by high phonetic ambiguity which could be decreased significantly by taking into consideration values of the adjacent vectors. In this case very similar /or even identical/ vectors can acquire different labels while quite different ones can get identical labeling depending on context information.

To simulate context-dependent interpretation of parameter vectors performed by the experts we have introduced the notions of centers of PSS and their periphery considering centers to be the most prominent and distinct vectors to interpret them in terms of AC-I. At first the sound string is analysed for centers which later guide the search for boundaries of PSS. The detected centers of the PSS and their readings in terms of AC-I serve as a context for the interpretation of their environments. It means that the reading given to the center is extended to all adjacent vectors /both preceding and subsequent/ untill they are consistent with the preconditions of the periphery-detecting rules which are significantly weaker than those of the center-detecting rules.

The whole procedure results in extracting PSS, which are later combined into segments corresponding to phonemes, as it was described above. The search for sonorants within vocal PSS is also organized as center-guided process. This enabled us to design various procedures for sonorant identification in pre-center, inter-center and post-center positions.

The algorithm is hierarchically organized according to confidence of inferences of the procedures included. The more confident procedures are activated earlier, narrowing the search area for succeding procedures.

Our work on the algorithm proved the spectrogram-reading approach to be very promissing and productive in view of speech recognition.

REFERENCES

/I/ R.A.Cole, A.I.Rudnicky, V.W.Zue, D.R.Reddy, "Speech as patterns on paper", in Perception and Production of Fluent Speech, R.A. Cole, ed., Lawrence Erl. Ass.,1980, pp.3-50.

/2/ M.A.Bush, G.E.Kopec, V.W.Zue, "Selecting acoustic features for stop consonant identification", Proc. ICASSP-83, pp. 742-745.

/3/ V.W.Zue, L.F.Lamel, "An expert spectrogram reader: a knowledge-based approach to speech recognition", Proc. ICASSP-86, pp. 23.2.I-4.

/4/ N.Carbonell, D.Fohr, et al., "An expert system for the automatic reading of French spectrograms", Proc. ICASSP-84, pp. 42.8.I-4.

/5/ N.Carbonell, J.-P.Damestoy, et al.,"APHODEX, Design and implementation of an acoustic-phonetic decoding expert system", Proc. ICASSP-86, pp. 23.3.I-4.

/6/ M.Liberman, "On the role of phonetic structure in automatic speech recognition", Proc. X-th ICPhS: Plenary Sessions. Symposia. Utrecht, I983, pp. 3I5-3I8.

/7/ J.Johannsen, T.MacAllister, et al.,"A speech spectrogram expert", Proc. ICASSP-83, pp. 746-749.

/8/ P.D.Green, A.R.Wood, "A representational approach to knowledge-based acoustic-phonetic processing in speech recognition", Proc. ICASSP-86, pp. 23.4.I-4.

/9/ J.Caelen, N.Vigouroux, "Producing and organizing phonetic knowledge from acoustic facts in multi-level data", Proc. ICASSP-86, pp. 23.5.I-4.