KNOWLEDGE-BASED ACOUSTIC-PHONETIC DECODING OF SPEECH : A CASE-STUDY WITH THE APHODEX PROJECT

J.P. Haton

CRIN-CNRS / INRIA, Nancy

ABSTRACT

The APHODEX project aims at investigating the role of Artifical Intelligence knowledge-based reasoning techniques in the acoustic-phonetic decoding (APD) of continuous speech. This paper constitutes an evaluation of this project. It briefly presents the present state of the APHODEX system and concentrates of some issues of APD that were raised during the project regarding several aspects : segmentation, amount of knowledge necessary for APD, choice of a proper unit decoding strategy.

1. INTRODUCTION

The acoustic-phonetic decoding of speech consists in automatically mapping the semi-continuous acoustic speech wave into a set of predefined discrete linguistic units such as phones, phonemes, syllables, etc. This is a very difficult operation which constitutes a major level in automatic speech recognition, especially for continuous speech, multi-speaker operation [1]. Despite substantial advances the problem has not yet received a totally satisfactory solution. One reason for that might be that APD makes it necessary to take into account a large body of information : data, facts, knowledge, contexts, etc. Following this idea, we launched in 1984 the APHODEX project with the aim of investigating to which extent the knowledge-based methodology developed in Artificial Intelligence may be helpful in solving the problem of APD.

After a brief recall of the different approaches to APD proposed so far we summarize the main features of APHODEX. We then present the major issues in APD in light of the experience we gained during the project.

2. APPROACHES TO ACOUSTIC-PHONETIC DECODING

The task of APD is of crucial importance in analytic speech recognition since the overall performances of any sentence recognizer depends heavily upon the quality of the phonetic decoding. The role of the acoustic-phonetic decoding level in speech recognition is threefold :

- extraction of pertinent features and cues from the acoustic signal,
- segmentation of the speech continuum into phonetically meaningful units such as phones, phonemes, syllables, etc.,
- labeling of the segments with fine phonetic labels and/or gross phonetic classes.

These three different tasks are highly interrelated and must moreover interact with the other linguistic processing levels in order to come out with the best possible phonetic lattice.

APD was initially considered as a simple pattern recognition problem. But the actual size and difficulty of the task were then clearly identified, particularly in relation with the major importance of coarticulation and context effects and of speech variability.

Present approaches to APD belong to three main categories :

- *stochastic modelling* [2] : the problem of optimally matching an input utterance against every possible concatenation of phonetic units can be expressed in terms of stochastic processing, especially in the framework of hidden Markov models. Such models make it possible to capture in a statistical way the variability of speech by processing huge amount of data. They provide one of most efficient framework for multi-speaker APD ;
- *connectionist neural-like modelling* [3], [4] : neural networks are experiencing a new growth of interest in different fields of Artificial Intelligence, including APD. Basic models (multi-layer perceptrons, Boltzmann machines, etc.) have been adapted to speech requirements, especiallyfor taking into account the inherent temporal nature of the speech phenomenon. New models more closely related to neuro-biological data have also been proposed, e.g. phonotopic maps or cortical columns [5]. Neural networks have achieved good performances in APD and represent a promising approach both for phonetic labeling and for preprocessing of speech data ;
- *knowledg-based reasoning* [6], [7] : the use of knowledge-based reasoning techniques is an alternative to the two previous statistically based approaches to APD. The major difficulty lies in the elicitation and formalization of a proper body of knowledge. Such techniques are used in the APHODEX project that will now be described in more details.

3. OVERVIEW OF APHODEX [8]

3.1. Basic ideas and motivations

Phonetic decoding by reading speech spectrograms is typically a knowledge intensive activity during which an experienced phonetician conducts an explicit and contextual reasoning based on the knowledge and expertise he gained by experience [9]. It seems therefore fruitful to elicitate and formalize this knowledge through a close interaction with a phonetician and by using the methodology developed in Artificial Intelligence for the design of knowledge-based systems. This idea was the basis of the APHODEX project when we started it in 1984. We considered at that time that the conjunction of the knowledge of an experienced spectrogram reader, François Lonchamp, on one hand and of our know-how in automatic speech recognition and knowledge engineering might help progressing in APD. Our main motivation was to gain a better understanding of the process of phonetic decoding and underlying processes. Another motivation was to implement an experimental knowledge-based system capable of carrying out the phonetic decoding of continuous speech in a multi-speaker way. The present state of this system together with its performances will now be briefly described.

3.2. Knowledge and architecture

Thanks to an in-depth study of the activity of spectrogram reading by our phonetician (cf. section 4.2.) we gathered a large body of procedural and declarative knowledge. This knowledge was then implemented in the APHODEX system into two forms :

- *procedures* coded in several pre-processors that operate directly on the speech wave and perform a coarse segmentation into phonetic segments as well as a classification of segments into broad phonetic classes (vocalic, fricative, plosive and others). Performances obtained so far are about 95 % of correct segmentation in the best cases ;
- *production rules* which constitute the knowledge base of an expert system. The inference engine of this expert system carries out a reasoning similar to the one developed by a phonetician in order to label each segment on a phonemic basis and, if need be, to refine the broad classification done by the pre-processors.

Here is a typical example of rule :

IF Segment is Plosive AND Burst spectral maximum is between 3000 Hz and 4500 Hz AND Right context is /i/ ou /e/ THEN /k/.

It is worth noticing that most rules are contextual (for instance here, the right context of the segment to be labeled must be an unrounded front vowel, i.e. in French /i/ or /e/). That enables the inference engine to carry out a contextual reasoning and to propagate constraints (phonetic, phonological, etc.) throughout the process in order to finally come out with an optimal phoneme lattice. All the acoustic events mentioned in the rules (formant trajectories, burst features, friction, etc.) are extracted automatically from the speech signal by robust, speaker-independent procedures.

Experimental results show that APHODEX is capable of decoding a sentence pronounced by any unknown French male speaker with a mean accuracy of 70 %. This percentage will progressively increase when new rules are added to the knowledge base. Comparatively, several experiments carried out for English and for French have shown that an expert spectrogram

reader can reach as much as 85 % of correct labeling.

4. ISSUES IN ACOUSTIC-PHONETIC DECODING OF SPEECH

We will now present some important issues in APD and propose some elements of solution that we developed in the framework of the APHODEX project.

4.1. Segmentation of the speech wave

The continuity of the speech signal is a major difficulty of speech recognition. A segmentation is therefore necessary in order to extract units on which the labeling process will then work. The problem is two-fold :

- firstly choose one or several proper units which can be either of infra-phonemic level (phones) or of phoneme level, or else of supra-phonemic level (diphones, syllables, triplets),
- then implement a segmentation process based on the temporal evolution of acoustic-phonetic features that must yield a solution as valid and consistent as possible.

The examination of some errors made by the segmenter of APHODEX led us to propose a hierarchical multi-segmentation solution, based upon the strategy used by the human expert. This method consists in building up a multi-level segmental representation of a sentence (dendrogram) with the help of a spectral difference function. This structure is then pruned out by using acoustic cues in order to yield the final segmentation which might be unique (in non-ambiguous cases) or multiple. This pruning is carried out in close interaction with the rule-based reasoning process.

4.2. Data and knowledge gathering

As stated previously the phonetic decoding of a sentence necessitates a large amount of knowledge of various types : articulatory, acoustic, phonetic, phonological, etc. This knowledge can be implicitly integrated in a system by the examination of huge amounts of speech data, as in stochastic or connectionist models. Despite the good performances obtained by such models, it seems nevertheless necessary to design some model for the explicit storage of knowledge. It seems that a knowledge-based reasoning APD can be more easily interfaced with other processing levels of a speech understanding system (for instance the feedback from the lexical level to the phonetic decoding). Moreover, this solution provides a convenient framework for gathering all available pieces of data and knowledge related to APD (the expert knowledge involved in spectrogram reading being only one aspect). The tools and methodology provided by artificial intelligence makes it easier to incrementally build up a kind of «collective memory» of APD for a given language. That constitutes a necessity for the future of research on speech communication.

4.3. Choice of a processing unit

The choice of a processing unit, for segmentation and for labeling, is of primary importance in the design of an APD system. As a matter of fact several units can be used at different steps of the process. The present version of APHODEX is based on a phoneme-like unit. This choice was motivated by the fact that the phonetician uses this unit through out his activity of spectrogram reading. Another feature interesting on a practical point of view is the limited number of phonemes which are necessary for the description of a language. However a phonemic unit presents serious drawbacks for APD, especially due to the context dependancy of phonemes that necessitates a very large number of rules to take into account the various contexts in which a phoneme has to be identified. That led us to adopt another processing unit, the phonetic triplet which an be defined as a phone with its phonetic contexct [10]. A large amount of work is still to be done in order to collect a set of triplets representative of the language but we nevertheless consider this units as a good compromise for APD.

4.4. Decoding control strategy

The APD reasoning operation must be controlled by an efficient strategy in order to avoid unnecessary hypotheses and to focus on relevant data. We developed in APHODEX a strategy inspired from the observation of the spectrogram reader who operates in two successive steps (cf. § 3.2.) A majority of APD systems use only a bottom-up strategy (from the acoustic data to phoneme labels). However a top-down strategy is also useful in order to verify hypotheses or to interact with the linguistic levels. In APHODEX the two types of control are used concurrently in an opportunistic manner. More work is still needed in order to design more sophisticated strategies similar to those used by an expert in difficult or ambiguous cases.

It is often necessary to make assumptions about the phonetic content of an utterance and to emit alternative, competing hypotheses about the succession of sounds. That corresponds to a kind of hypothetical reasoning for which specific techniques have been developed in AI in order to maintain the overall truth and consistency of the deductions made during the decoding. We are implementing hypothetical reasoning in APHODEX, based on various types of knowledge including phonological variations. This method gives substantial improvements in the decoding accuracy, especially when there is some ambiguity or when contextual effects are important.

Two important lessons drawn from the examination of spectrogram reading activity concern the strategy of decoding. The first one consists in systematically relying phonetic labeling decisions to several acoustic features rather than a single one. The second can be called delayed decision strategy since it consists in postponing decisions until enough evidence has been accumultaed in favor of a particular label.

5. CONCLUSION

This paper has dealt with some aspects of a major problem in automatic speech recognition, i.e. acoustic-phonetic decoding of continuous speech. We have especially presented the usefulness of Artifical Intelligence knowledge-based techniques in this area and discussed important issues in the light of the APHODEX project developed in our group.

Despite the very good performances obtained so far in APD by statistical models, we consider that knowledge-based techniques have some usefulness both for gathering relevant data and knowledge and for implemnting practical systems. An explicit knowledge-based reasoning in APD also makes it easier to implement feed back links from linguistic processing levels to APD.

6. REFERENCES

[1] KLATT, D. (1977), "Review of the ARPA Speech Understanding Project", *JASA*, 62, pp. 1345-1366.

[2] SCHWARTZ, R.M. et al. (1984), "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition", *Proc. Int. Conf. ASSP*.

[3] BOURLARD, H., WELLEKENS, C.J. (1989), "Speech Dynamics and Recurrent Neural Networks", *Proc. ICASSP-89*, Glasgow.

[4] WAIBEL, A., SAWAI, H., SHIKANO, K. (1989), "Consonant Recognition by Modular Construction of Large Phonemic Time Delay Neural Networks", *Proc. ICASSP-89*, Glasgow.

[5] GUYOT, F., ALEXANDRE, F., HATON, J.P. (1989), "Toward a Continous Model of the Cortical Column : Application to Speech Recognition", *Proc. ICASSP-89*, Glasgow. [6] GREEN, P.D. et al. (1987), "A Speech Recognition Strategy Based on Making Acoustic Evidence and Phonetic Knowledge Explicit", *Proc. European Conf. Speech Technology*, Edinburgh.

[7] MEMMI, D., ESKENAZI, M., MARIANI, J., NGUYEN-XUAN, A. (1983), "Un système expert pour la lecture de sonagrammes", *Speech Com.*, vol. 2, n° 2-3, pp. 234-236.

[8] CARBONELL, N., FOHR, D., HATON, J.P. (1987), "An Acoustic-phonetic Decoding Expert System", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 1, n° 2, pp. 207-222.

[9] ZUE, V.W., COLE, R.A. (1979), "Experiments on Spectrogram Reading", *IEEE-ICASSP*, Washington, D.C., pp. 116-119.

[10] LAPRIE Y., HATON, J.P., PIERREL, J.M. (1990), "Phonetic Triplets in Knowledge-based Approach of Acoustic-phonetic Decoding", *Proc. Conf. Speech and Language Processing*, Kobé, Japan.

Acknowledgment :

The author gratefully acknowledges the contribution of members of the RFIA group to the APHODEX project : Anne Bonneau, Noëlle Carbonell, François Charpillet, Jean-Paul Damestoy, Mahieddine Djoudi, Dominique Fohr, Dominique François, Ramez Hajislam, Marie-Christine Haton, Yves Laprie, Jean-Marie Pierrel.