AUTOMATIC LABELLING OF SPEECH SIGNAL INTO PHONETIC EVENTS

H. Kabré, G. Pérennou and N. Vigouroux

Institut de Recherche en Informatique
Toulouse, France

ABSTRACT

In this paper, we give the general principles behind an automatic system, developed at IRIT Lab, and capable of labelling speech signal for phonetic events. When using this sytem, the results secured on English, French and Swedish corpora demonstrate that the labelling operation becomes completely independent from either language, corpus or speaker. Moreover, this operation requires no manual adaptation or training whatsoever.

1. INTRODUCTION

Automatic labelling of speech corpora is an increasingly important problem, when considering present-day development of recorded speech databases —e.g., the DARPA Project ones.
In Europe, within the scope of the SAM ESPRIT Project —involving this kind of databases for multilingual corpora— the question has quickly arisen both as to how to adapt these various automatic labellers to different languages, and as to how to process speech material without having to resort either to a manual adaptation or to some kind of language, speaker or corpus training.
The latter problem is the one considered, here, as we are presenting SAPHO —the phonetic front-end of our automatic labeller.

2. SYSTEM STRUCTURE

Our labelling system procedes in two successive main stages:
- in the course of the first one, the signal is both segmented and labelled into phonetic events (SAPHO component);
- in the second stage, these events become aligned onto a phonetic transcription, supplied beforehand by a phonetician (VERIPHONE component).

A more detailed description of this can be found in [2], as only the general principles triggering decision are reported here.
1) The automatic alignment, arrived at, does not require any fine characterization of the phonetic events involved; macro-class labels being quite sufficient to do the job (As an extreme example, if the sequence [tom] were to be aligned onto the five-event string [+occl] [+fri] [+vow] [+voc] [+occ], the phoneme [t] would easely be identified as made up of the first two elements of the string, while [o] would readely be identified as the third element and [m] as the last two elements).
2) Under such conditions, the whole set of permanent acoustic parameters, necessary for recognition, does not have to be used up.
3) An appropriate, limited choice, among these, will bear upon the minimum subset of the most robust parameters, given the target set for the exercise; i.e., labelling that is independent from either language, speaker or corpus.
In [4], various parameters can be found, which were used in automatic labelling. Some of them describe signal amplitude, others spectrum.
We chose two parameters that are universally used by manual labellers; namely, the amplitude parameter and the zero crossing rate —both of wich contain enough information to accomplish the contemplated task.
Both afford the advantage of an identical performance over whole sets of languages, speakers and corpora.

This work is supported by the SAM ESPRIT Project and the GDR-PRC Communication Homme-Machine Program.

These parameters have to be pre-processed in order to achieve an optimal automatic segmentation.
In Fig.1, for example, three parameter forms can be seen to characterize amplitude.
In 1c, maximun amplitude is evaluated over each one of the 4ms successive frames; this amplitude has undergone a non-linear smoothing (NLSA parameter) that does preserve major instances of signal discontinuity.
In 1d, mean amplitude (energy) over each 8ms frame is evaluated.
These two amplitude values can be directly compared to the initial waveform given in 1b.
Our choice went to the logarithmic NLSA amplitude —normalized LNLSA.
This normalization occurs at two different levels:
- with respect to the whole corpus, in order for the amplitude, thus normalized, to vary within the [0, 1] interval,
- with respect to a local signal interval, by taking the ratio of amplitude to maximum amplitude within a ± 0.25 ms window that is centered upon the instant considered (whenever this maximum amplitude falls below a floor value, the latter is taken, instead, as the denominator of the ratio).

This representation of amplitude is advantageous in at least three ways :
- it preserves essential contrasts between successive phonemes (the NLSA parameter can be compared to the mean amplitude one; allowing to observe that, with the latter, the contrast "closed/nasal consonant" is all but lost, whereas it comes out enhanced with NLSA);
- amplitude comes out smooth, while essential instances of discontinuity are preserved;
- amplitude is strongly correlated to phoneme aperture; the effects of the mean sound intensity variation, within the phrase, being attenuated by local normalization.

Similar treatments are applied to the zero crossing rate given in 1e; although, in this case, normalization is global over the whole corpus.
The parameter thus obtained is LNLSZ.
Both parameters, LNLSA amplitude and LNLSZ zero crossing rate, constitute the starting basis for an evaluation of cues, enabling to label each 4ms within one of the categories appearing in the table on Fig.2.

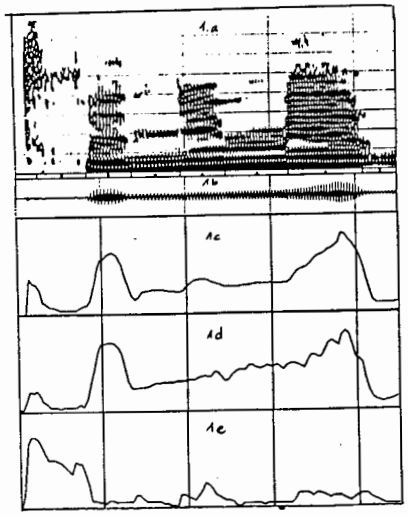

Fig.1 - Temporal parameters : 1a) Spectrogram, 1b) waveform, 1c) NLSA normalized amplitude, 1d) 8ms mean amplitude, 1e) NLSZ zero crossing rate.

Events are obtained, next, when regrouping frames by labels of equal values, while smoothing off any segment that is too short.

K	strong syll	S	s like fric
W	weak syll	C	short S
L	acute voc	Z	z like fric
U	grave voc	F	weak fric
O	voiced occl	X	x like fric
Q	unvoiced occl		

Fig.2 - Table 1 : Phonetic events.

3. EVALUATION

The quality of various events, obtained, can be evaluated thanks to the different kinds of results.
The ones given in Table 2, Fig.3, show quality of automatic segmentation, as this compares to manual labelling, when the latter is obtained over EUROM-0 French, English and Swedish corpora, without any kind of either manual adaptation(-phase) or learning.

These results remain very steady from one language to the next. Furthermore, boundary accuracy is of the same order as what is observed when comparing between 'handlabellers' performances.

Pass	Language	Phoneme Number	Speaker Number	Surseg. Ratio	QC(*) ±13 ms	QC ±17 ms	QC ±21 ms	QC ±25 ms
1	English	4476	4	2.10	0.850	0.890	0.914	0.933
2				1.44	0.794	0.843	0.872	0.888
3				1.27	0.763	0.815	0.846	0.865
1	French	2909	2	2.74	0.866	0.918	0.938	0.950
2				1.78	0.823	0.894	0.919	0.933
3				1.53	0.793	0.870	0.898	0.913
1	Swedish	1379	1	2.15	0.782	0.833	0.861	0.884
2				1.58	0.742	0.793	0.826	0.846
3				1.35	0.700	0.762	0.803	0.822

(*)QC = [n / N] , N = number of manual boundaries, n = those ones that have an approximation to an automatic boundary less than ±x ms (x=13, 17, 21, 25).

Fig.3 - Table 2 : Quality of the SAPHO segmentation.

The other results also display great steadiness, both over various corpora and from one language to the next.
The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process.
Modelizing a given phonetic unit, belonging to a given language, boils down, therefore, to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units. In the way of phonetic units, it is of course much better to choose contextual allophones, for a more homogeneous spread of the various realizations.
In addition to this process —which is likely to occur in every language— there are properties —also common to all languages— such as the presence of events that are specific to natural classes of phonetic units.
This is illustrated in the table on Fig.3, where stops can be seen generally to entail an event Q. It is clear, however, that in this respect languages differ from each other through their respective phonological systems, and that the stochastic laws pertaining to various phonetic units must be specifically estimated for each such unit.

ENGLISH

ph	K	W	U	L	O	Q	Z	F	X	C	S	ssr
p	0	0	0	1	2	0	6	6	6	0	0	1.2
t	0	0	0	0	0	39	35	12	8	6	40	2
k	0	0	0	5	0	6	14	19	13	0	6	1.6

FRENCH

ph	K	W	U	L	O	Q	Z	F	X	C	S	ssr
p	0	0	0	10	8	96	0	14	7	0	1	1.4
t	0	1	11	11	3	90	8	12	3	5	12	1.6
k	0	2	18	6	4	104	24	20	14	0	8	2

Fig.4 - Average number of phonetic events in [p] [t] [k] phonetic units and sursegmentation rate (ssr) for English and French.

Thus, tables on Fig.4 shows that, in English, [t] becomes realized often (probability in the order of 40%) as Q+S. This combination does occur in French, as well, but with a lesser frequency (ca. 12 % prob.). Conversely, the combination of Q with a vocalic segment (W, L or U) seldom occurs in English, whereas it is frequent in French (ca. 23 %).

The results, presented in this paper, show that segments, provided by a handlabeller in order to account for a realization of phonetic units showing up in a transcription, generally are compounds that can otherwise be broken down into a set of a few phonetic elements made available by the SAPHO automatic process.
Modelizing a given phonetic unit, belonging to a given language, boils down therefore to specifying the stochastic laws which pertain to it and which steer a combination of events leading up to a realization of these units.
In addition to this process —which is likely to occur in every language— there are properties —also common to all languages— such as the presence of events that are specific to natural classes of phonetic units. This is illustrated in the table on Fig.4, where stops can be seen generally to entail an event Q.

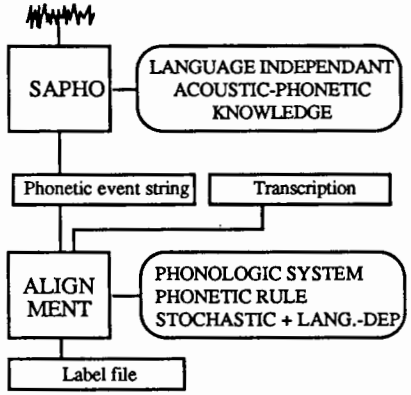

Fig.5 - The two levels of automatic labelling; language modelling implications.

It is clear (Fig.5), however, that in this respect languages differ from each other through their respective phonological systems, and that the chance laws pertaining to various phonetic units must be specifically estimated for each such unit. Thus, the table on Fig.4 shows that, in English, the phoneme [t] becomes realized quite often (probability in the order of 40%) as Q+S. This combination does occur in French, as well, but with a lesser frequency (ca. 12 % prob.). Conversely, the combination of Q with a vocalic segment (W, L or U) seldom occurs in English, whereas it is frequent in French (ca. 23 %).

4. CONCLUSION

The results we have secured over English, French and Swedish speech corpora, demonstrate the feasibility of labelling phonetic events that are language-, speaker-, as well as corpus-independent. However, these results should be reinforced both over larger corpora and over a more numerous set of languages. The results, presented here, were secured with the SAPHO System, which makes use of information relating only to amplitude and zero crossing rate.
We are now working at an efficient use of these events in automatic alignment and on pre-selection of sub-vocabularies within large lexicons.

The authors are thankful to Prof. J. F. Malet, CSU Sacramento, for this prompt translation of their original French ms.

5. REFERENCES

[1] C.J.M. Hoeckel, "The Reliability of Manual Labelling of Continuous Speech", Proceeding of ESCA Tutorial Day and Workshop on Speech Input/ Output Assessment and Speech Databases, paper 5.5.1-5.5.4.
[2] G. Pérennou, M. de Calmès, J.M. Pécatte, N. Vigouroux, "Phonetic-String Alignment for an Automatic Labelling of Speech Corpora", in Proceedings of Workshop on Speech Input/Output Assessment and Speech Databases, The Netherlands, 20-23 September, pp. 5.4.1-5.4.4.
[3] S. Seneff, V. Zue "Transcription and Alignment of the TIMIT Database," in Getting Started With The DARPA TIMIT CD-ROM.
[4] ESPRIT Project 2589, Intermediate Report, 1st March 1989 - 28 February 1990.
[5] V.W. Zue, R.M. Schwartz, "Acoustic Processing and Phonetic Analysis", in Trends in Speech Recognition W.A. Lea (ed.), pp. 101-124.l.