A PRELIMINARY STATISTICAL EVALUATION OF MANUAL AND AUTOMATIC SEGMENTATION DISCREPANCIES

Piero Cosi (*), Daniele Falavigna (**) and Maurizio Omologo (**)

(*) Centro di Studio per le Ricerche di Fonetica - C.N.R., P.zza Salvemini, 13, 35131-Padova (Italy)

(**) IRST, Istituto per la Ricerca Scientifica e Tecnologica, Panté di Povo, 38050-Trento (Italy)

ABSTRACT

The accuracy of automatic alignment systems will always be checked using references manually segmented by phonetic or speech communication experts. Consequently, the statistical characterisation of manual segmentation discrepancies becomes quite an important issue. To achieve the goal of finding a better statistical description of these discrepancies, two different experiments were carried out. The first pilot experiment, which was focused on the segmentation of a small subset of an Italian isolated-word speech data-base, was carried out only to better define the second experiment regarding the segmentation of few continuous Italian sentences.

This work has been developed in part under the ESPRIT project "Speech Assessment Methodology" (SAM).

1. INTRODUCTION

Phonetic or phonemic labelling of speech signals is normally performed manually by phonetic or speech communication experts. Even if various attractive graphic and acoustic tools were simultaneously available, there will always be some disagreement among skilled human labelling experts in the results of labelling the same waveform. In fact, due to human variability of visual and acoustic perceptual capabilities and to the difficulty in finding a clear common labelling strategy, the manual labelling procedure is implicitly incoherent. Another important drawback of manual intervention in labelling speech signals is that it is extremely time consuming. Considering these and other disadvantages, the development of methods for automatic labelling of speech data is becoming increasingly important [1],[2],[3],[4],[5]. In fact, automatic labelling systems minimise assessment time of input/output speech data-bases and are at least implicitly coherent. In other words, obviously using the same strategy, "if they make some errors they always make them in a coherent way". Moreover, even if segmentation and labelling are avoided by some recently developed successful automatic speech recognition systems, generally based on Hidden Markov Model techniques, a completely labelled true continuous speech database will always be of interest for linguistic and phonetic research.

In this work two experiments were carried out, but the first one was intended only to bootstrap the second. In fact, to become familiar with the segmentation environment, the phonetic alphabet chosen and the common segmentation strategy, a subset of the IRST-CSRF IWSDB isolated-word speech database [6] was first segmented by four experts and their segmentation discrepancies were analysed. Successively, three experts were considered for the second experiment regarding the segmentation of a particular subset of the IRST-MAIA CSDB continuous speech database [7], consisting of ten sentences. Finally, an automatic segmentation algorithm developed at IRST [8] was applied to the same corpus and the automatic boundary positions obtained were compared to those previously imposed by humans. In this paper the two experiments will be described as well as the statistical results, and conclusions on the use of manual versus automatic segmentation will be drawn.

2. "BOOTSTRAPPING" EXPERIMENT

An experiment was created with the aim of familiarising the human experts with the following three aspects of the segmentation procedure:

a) the segmentation environment, in other words the audio and visual facilities available for segmenting the speech material;
b) the chosen phonetic alphabet;
c) the common adopted segmentation strategy.

The speech analysis and visualisation system called PTS, developed within the ESPRIT-2589 Project, (SAM) Multi-Lingual Speech Input/Output Assessment, Methodology and Standardisation [9], was used to segment the speech material. The Italian SAMPA phonetic alphabet [9] was adopted, and the following common segmentation strategy interactively developed.

2.1 SEGMENTATION STRATEGY

Three principal rules must be applied:

RULE 1)
set the environment parameters to the same values (e.g. PTS visual and audio parameters, D/A specific parameters, etc..);
RULE 2)
indicate the boundaries between phonetic units that are unambiguously detected by visual and audio inspection of the speech material;
RULE 3)
for each of the ambiguous cases, play "windows" placed on the left and on the right of the hypothesised boundary position, and iteratively extended, until the second and the first unit, respectively, are perceived. These two positions, which are generally different, and the aim of which is to delimit the speech area within which the final boundary should reside, are memorised. Finally the target boundary should be set at the clearest acoustic event near the middle point of the above found speech area. Figure 1 graphically illustrates this procedure.

Various practical suggestions also were given to the experts.

2.2 PRACTICAL SUGGESTIONS

a) always listen to speech segments never shorter than the mean length of a word (4 to 5 phonetic units);
b) utilise spectral information (sonogram) only in cases of great uncertainty (e.g. for diphthongs, hiatuses or nasal followed by liquid,.....);

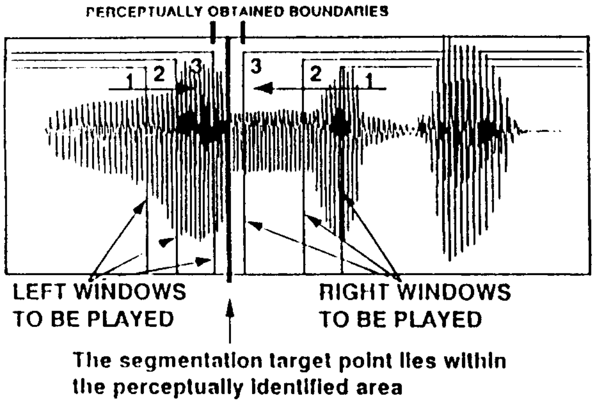
PERCEPTUALLY OBTAINED BOUNDARIES
1
2
3
3
2
1
LEFT WINDOWS
TO BE PLAYED
RIGHT WINDOWS
TO BE PLAYED
The segmentation target point lies within
the perceptually identified area

Fig. 1. Graphic illustration of segmentation RULE 3 (see text).

c) always set the target boundary in correspondence to a significant event visually discovered in the speech waveform (e.g. the first zero crossing of the speech waveform in the first pitch period of a voiced sound following an unvoiced sound);

d) for diphthongs, both ascending and descending, assign to the weak vowel (/i/,/u/) almost 1/3 of the whole diphthong duration.

e) as for the so called "epenthetic silence" characterised by distinct regions of weak energy separating sounds that involve a change in voicing (e.g. fricative followed by semiconsonant or nasal), set the target boundary at the end of this phenomenon, consequently assigning this region to the first of the two phonetic units involved.

2.3 SPEECH MATERIAL

The IRST-DB isolated-word speech database [6] was used in this experiment initially to facilitate segmentation. The whole DB comprises five pronunciations of 94 phonetically significant stimuli produced by 10 male and 10 female speakers and two pronunciations of the same list of stimuli produced by other 15 male and 15 female speakers, for a total of 15040 stimuli, and all the speech material was digitised at 16kHz sampling frequency. The manual segmentation of a complete list of 94 stimuli of the same randomly chosen speaker was carried out by four experts and their segmentation boundaries were analysed.

2.4 STATISTICAL ANALYSIS OF SEGMENTATION DECISIONS

The ELSA software [10], developed within the above mentioned SAM Project, was used to describe statistically the discrepancies of the four different manual segmentations. Tables 1 and 2a-b summarise the results.

	PLO	AFF·	FRI	NAS	LIQ	SEM	VOW	ALL
PLO			9		3	5	3	3
AFF							5	5
FRI	4						4	4
NAS	8	14	6				5	6
LIQ	4	6	3	7			5	5
SEM							10	10
VOW	13	3	8	7	7	9	6	11
ALL	7	5	5	7	6	6	5	7

Table 1. Manual segmentation mean deviations (ms) for the isolatεd-word experiment.

The accuracy of manual segmentations was computed comparing three test manual alignment settings with the fourth one, which was considered the reference. Test and reference settings were circulated four times to cover all the possible combinations. In Table 1 the mean deviations of segmentation boundaries are referred to phonetic units grouped together in phonetic classes to better group the results and to give better evidence of possible common tendencies of alignment errors. The classes on the left refer to the first item while those on the top refer to the second item of the phonetic transition.

(a)	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO	0/0	0/0	10/12	0/0	48/48	32/36	343/348	433/444
AFF	0/0	0/0	0/0	0/0	0/0	0/0	73/78	73/78
FRI	40/42	0/0	0/0	0/0	0/0	0/0	156/156	202/204
NAS	38/42	4/6	6/6	0/0	0/0	0/0	112/114	169/180
LIQ	6/6	6/6	6/6	17/18	0/0	0/0	158/162	202/210
SEM	0/0	0/0	0/0	0/0	0/0	0/0	35/42	35/42
VOW	146/174	12/12	76,84	113/120	133/138	6/6	26/30	851/1098
ALL	407/444	76/78	191,204	172/180	202/210	38/42	1065/1098	2514/2820

(b)	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO			55-95		93-100	75-96	97-99	96-99
AFF							86-97	86-97
FRI	84-99						98-100	96-100
NAS	78-96	30-90	61-100				94-100	89-97
LIQ	61-100	61-100	61-100	74-99			94-99	93-98
SEM							69-92	69-92
VOW	78-89	76-100	82-95	88-97	92-98	61-100	70-95	75-80
ALL	89-94	91-99	89-96	91-98	93-98	78-96	96-98	88-90

Table 2. Manual segmentation (a) absolute [n/N] and (b) percentage [%] correct values for the isolated-word experiment.

In Tables 2a-b, with the same grouping, the correct segmentation absolute values and percentages are shown respectively, with an error criterium of $\pm$20ms. Single percentage values are associated with a statistical measure that was introduced because of the poor coverage of certain phonetic classes. A minimum and maximum percentage of correct segmentation, which covers the target value with a probability of 95%, is indicated (with a bar). As indicated in Table 1, considering all phonemes followed by all phonemes (lower right value), a 7ms mean deviation has been observed, while the transition between nasals and affricates (14ms) or vowels and plosives (13ms) or semiconsonants and vowels (10ms) gives the highest mean deviations. The transitions between semiconsonants and vowels together with those of vowels and plosives give the worst segmentation agreement among the experts (see Table 2). The rather low global result (lower right value, 88-90%) clearly emphasises the need to find a simpler and more uniform segmentation strategy and also the need of a certain period of training, even for human experts skilled in phonetics or speech communication, to make the segmentation results more reliable. The analysis of these results together with a more detailed phoneme by phoneme inspection of expert disagreements highlighted which were the "difficult" phonemes or classes of phonemes to segment and consequently has allowed expert attention to focus on them, for further experiments. Through various refinement steps the segmentation criteria were interactively discussed in order to better define an unambiguous strategy.

3. CONTINUOUS SPEECH SEGMENTATION EXPERIMENT

With the previously defined and tuned experimental setting, a second experiment regarding the segmentation of 10 continuous sentences was considered. Only three out of four experts were involved in this experiment. The IRST-MAIA CSDB continuous speech data-base [7] was used. It consists of sentences of the type,

Mi hanno detto di andare alla fotocopiatrice vicino alla segreteria ORI
(They told me to go to the Xerox-machine near the ORI secretariat)

recorded in a quiet office room.

Segmentation mean deviations and absolute values and percentages of correct segmentation, with an error criterium of +20ms, are respectively illustrated in Tables 3 and 4a-b, while in Figure 2 the percentage correct values of manual settings for all the phonemes followed by phonemes belonging to the particular phonetic classes are illustrated for various error criteria (+5, +10, +20, +30, +40ms). From the "mean deviations" point of view (Table 3), semiconsonants and vowels still present difficulty for segmentation. Transitions between vowels and plosives were better segmented in this obviously more difficult continuous-speech case than in the previous isolated-word case, and this perhaps is due to more emphasis being given to these particular classes after the first bootstrapping experiment. For the transitions between vowels and affricates or vowels and vowels the opposite situation has been observed. In fact, their mean deviation was raised from 3ms to 14ms and from 6ms to 14ms respectively, thus indicating an obvious increment in the difficulty of segmentation, passing from the isolated-word to the continuous-speech case.

	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO					7	2	3	3
AFF						3	5	3
FRI	3						3	3
NAS	7		5				4	5
LIQ			9	6			5	5
SEM							16	16
VOW	7	14	6	8	6		14	8
ALL	6	14	5	8	6	3	5	6

Table 3. Manual segmentation mean deviations (ms) for the continuous-speech experiment.

The global performance ("all-phonemes" class followed by "all-phonemes" class) obtained in this experiment, as for the absolute values as for the percentages of correct segmentation, are indicated by the lower right values of Table 4. They indicate a slight increment in the performance of the experts, in comparison with those of the previous experiment. Even if the phonetic coverage was very different, especially for certain classes, from that of the isolated-word case, thus contributing to explain the differences in the performance of the two experiments, the adopted "focus of attention" strategy, following the observations made after the first bootstrapping experiment, seems to have augmented the robustness of the common segmentation criteria making them less ambiguous.

4. MANUAL vs AUTOMATIC SEGMENTATION

In order to verify the possibility of adopting an automatic segmentation algorithm for speech database assessment the IRST alignment system [8] was applied to the same continuous speech corpus utilised in the previous experiment and the results were compared with those previously obtained by human experts.

4.1 AUTOMATIC ALIGNMENT SYSTEM

The algorithm that was used for automatic labelling is described in detail in [8]. In what follows only the main issues are summarised. A preliminary step is required to define a reference set of LPC unit prototypes. These units represent each phonetic label that can be produced by the preliminary grapheme-to-phoneme transcription. The algorithm consists of the following steps:

a) a grapheme to phoneme conversion is applied to the orthographic representation of the utterance to be labelled;
b) LPC analysis is carried out on the corresponding speech waveform;
c) a spectral variation function [11] is computed to emphasise significant acoustic changes in the speech signal;
d) a constrained DTW(Dynamic-Time-Warping)-based algorithm aligns the speech waveform with a concatenation of the LPC unit prototypes corresponding to the graphemic representation of the given utterance;
e) the local energy contour is used to improve the resulting boundary positions.

(a)

	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO	0/0	0/0	0/0	0/0	18/18	9/9	91/93	118/120
AFF	0/0	0/0	0/0	0/0	0/0	3/3	15/15	18/18
FRI	6/6	0/0	0/0	0/0	0/0	0/0	57/57	63/63
NAS	28/30	0/0	9/9	0/0	0/0	0/0	38/39	75/78
LIQ	0/0	0/0	3/3	9/9	0/0	0/0	78/78	90/90
SEM	0/0	0/0	0/0	0/0	0/0	0/0	8/12	8/12
VOW	64/69	14/18	39/39	55/60	72/72	0/0	31/42	316/354
ALL	113/120	14/18	63/63	69/78	90/90	12/12	336/354	738/789

(b)

	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO					82-100	70-100	92-99	94-100
AFF						44-100	80-100	82-100
FRI	61-100						94-100	94-100
NAS	79-98		70-100				87-100	89-99
LIQ			44-100	70-100			95-100	96-100
SEM							39-86	39-86
VOW	84-97	55-91	91-100	82-96	95-100		59-85	86-92
ALL	88-97	55-91	94-100	80-94	96-100	76-100	92-97	92-95

Table 4. Manual segmentation (a) absolute [n/N] and (b) percentage [%] correct values for the continuous-speech experiment.

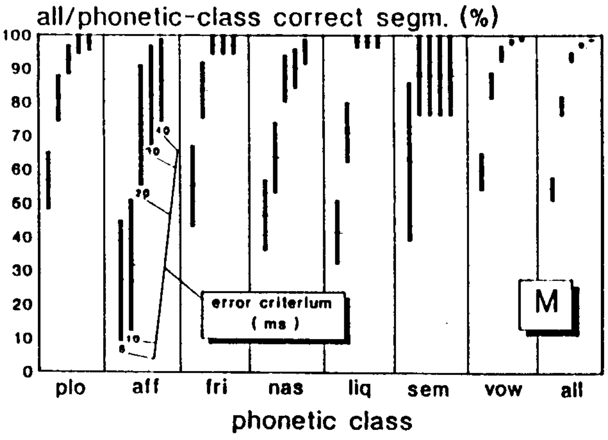

Figure 2. Correct segmentation percentage values of manual settings in the case of all phonemes followed by phonemes belonging to a particular phonetic class for various error criteria (+5, +10, +20, +30, +40ms) and for the continuous-speech experiment . Single percentage values are associated with a statistical measure which was introduced due to the poor coverage of certain phonetic classes (see text).

4.2 AUTOMATIC vs MANUAL PERFORMANCE

The results of the application of the automatic alignment system are given in Tables 5 and 6a-b. Both for the mean deviations (Table 5) and the absolute (Table 6a) or the percentage (Table 6b) values of correct segmentation, a big difference still exists between human experts and the chosen automatic system. The alignment system performance is expected to increase using more allophonic units and more training material, especially for the labelling of "true" continuous speech.

	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO					24	10	27	25
AFF						10	8	8
FRI	2						13	12
NAS	11		12				30	21
LIQ			19	44			41	41
SEM							38	38
VOW	23	9	10	27	50		43	30
ALL	16	9	18	30	45	10	29	27

Table 5. Automatic segmentation mean deviations (ms) for the continuous-speech experiment.

(a)

	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO	0/0	0/0	0/0	0/0	11/18	8/9	45/93	64/120
AFF	0/0	0/0	0/0	0/0	0/0	3/3	14/15	17/18
FRI	6/6	0/0	0/0	0/0	0/0	0/0	48/57	54/63
NAS	28/30	0/0	6/9	0/0	0/0	0/0	22/39	56/78
LIQ	0/0	0/0	2/3	5/9	0/0	0/0	40/78	47/90
SEM	0/0	0/0	0/0	0/0	0/0	0/0	3/12	3/12
VOW	41/69	15/18	33/39	41/60	30/72	0/0	9/42	198/354
ALL	90/120	16/18	47/63	50/78	41/90	11/12	196/354	479/789

(b)

	PLO	AFF	FRI	NAS	LIQ	SEM	VOW	ALL
PLO					39-80	56-98	38-58	44-62
AFF						44-100	70-99	74-99
FRI	61-100						73-91	75-92
NAS	79-98		35-88				41-71	61-81
LIQ			21-94	27-81			40-62	42-62
SEM							9-53	9-53
VOW	48-70	67-97	70-93	56-79	31-53		12-36	51-61
ALL	67-82	67-97	63-84	53-74	36-56	65-99	50-60	57-64

Table 6. Automatic segmentation (a) absolute [n/N] and (b) percentage [%] correct values for the continuous-speech experiment.

As for the manual segmentation case, Figure 3 shows the percentage of correct alignments of the automatic system, when all the phonemes are followed by phonemes belonging to particular phonetic classes, are illustrated for different error criteria (±5, ±10, ±20, ±30, ±40ms). Similar performance, but obtained on different corpora, by other ESPRIT automatic alignment systems has been reported in the literature [1], [2], [3].

5. CONCLUSIONS AND FUTURE WORK

The results of manual alignment comparisons indicate that some of the boundary discrepancies are, of course, due to the need for better phoneme segmentation strategies, but it is also clear that the improvement for some sounds must remain limited due to the implicit fuzziness of the criteria used for delimiting them, as for example in the case of semiconsonants and vowels in diphthongs.

At present, the comparison of segmentation performance obtained by a statistically well described group of experts and even very advanced automatic systems [5], indicates the actual superiority of human capabilities, even if the behaviour of manual labellers in a bigger task, as that considered in [5], has to be analysed in the future in order to confirm this very preliminary conclusion. In spite of that the applicability of semi-automatic systems, with interactive human intervention, for speech-database assessment has already become a reality [4].

The statistics of human and automatic alignment discrepancies can obviously be utilised to reduce the degree of ambiguity of manual segmentation criteria and also to test new experts. Moreover, these statistics can improve the automatic systems forcing them to hardly concentrate in those cases badly analysed in comparison with humans.

Further research will be carried out, in the future, to better describe statistically human discrepancies in "fluent" continuous speech segmentation, a clearly more difficult task than an isolated-word or "controlled" continuous speech case. Particular attention will be given to the phonetic coverage of future speech data-bases. Parallely an improvement of the present automatic alignment system [8] will be looked for, obviously utilising preliminary results obtained in this work.

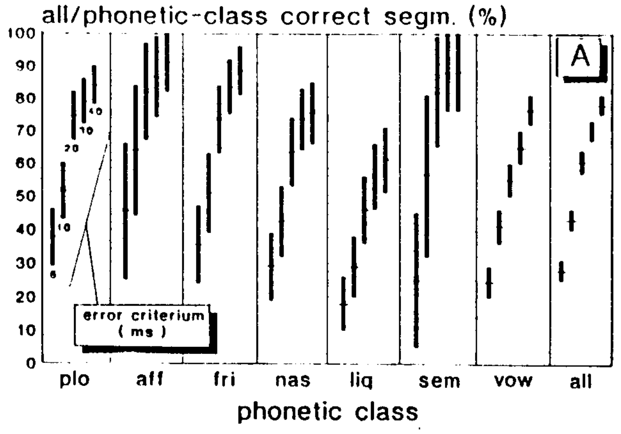

Figure 3. Correct segmentation percentage values of automatic settings in the case of all phonemes followed by phonemes belonging to a particular phonetic class for various error criteria (±5, ±10, ±20, ±30, ±40ms) and for the continuous-speech experiment.

REFERENCES

[1] P. Dalsgaard,"Semi-Automatic Phonemic Labelling of Speech Data using a Self-Organising Neural Network ", Proceedings of EUROSPEECH 89, September 26-28, 1989, pp. 541-544.

[2] C. Dours, M. de Calms, H. Kabr, J.M. Pcatte, G. Prennou and M. Vigoroux, "A Multi-Level Automatic Segmentation System: SAPHO and VERIPHONE ", Proceedings of EUROSPEECH-89, Paris France, September 1989, Vol. 2, pp. 83-87.

[3] T. Svendsen and K. Vale, "Automatic Alignment of Phonemic Labels with Continuous Speech" , Proceedings of ICSLP-90, Kobe Japan, November 1990, Vol. 2, pp. 997-1000.

[4] H.C. Leung and V.W Zue, "A procedure for Automatic alignment of Phonetic Transcription with Continuous Speech", Proceedings of ICASSP-84, March 1984, pp. 2.7.1-2.7.4.

[5] A. Ljolje and M.D. Riley, "Automatic Segmentation and Labelling of Speech", Proceedings of ICASSP-91, Toronto, Canada, May 1991.

[6] D. Falavigna and M. Omologo, "IRST-CSRF Data Base ", IRST Internal Report 1989.

[7] M. Omologo, "Sintesi del segnale vocale per il progetto MAIA: procedure di estensione del vocabolario e definizione di un modulo prosodico ", IRST Internal Report #9012-14,1990.

[8] D. Falavigna and M. Omologo, "A DTW Based Approach to the Automatic Labelling of Speech According to the Phonetic Transcription", Proc. Eusipco '90, Vol. 2, pp. 1139-1142.

[9] A.J. Fourcin, G. Harland, W. Barry and W. Hazan eds., "Speech Input and Output Assessment, Multilingual Methods and Standards ", Ellis Horwood Books in Information Technology, 1989.

[10] C. Bourjot, A. Boyer and D. Fohr, "Semi Automatic Labelling Assessment Software" SAM-ESPRIT-Document.

[11] M. Omologo and D. Falavigna, " A spectral Variation Function for Acoustic Speech Segmentation", Proc. VERBA '90, pp. 365-372.