NON-LINEAR ANNOTATION OF MULTI-CHANNEL SPEECH DATA

W.J. Hardcastle *, A. Marchal +, K. Nicolaidis *, N. Nguyen-Trong +

* University of Reading, UK., + CNRS, Aix-en-Provence, France

ABSTRACT - The principles of the non-linear annotation used in the ACCOR project are described with reference to connected speech data selected from the EUR-ACCOR multi-channel speech database.

INTRODUCTION

Segmentation and labelling of the speech signal is problematic since there is no one-to-one correspondence between the physical and the linguistic levels of representation: boundaries between "neighbouring" segments (as defined from a symbolic transcription) are blurred by the speech encoding process. Various approaches have been proposed attempting to relate the symbolic levels (syntactic, phonemic, phonetic etc.) to the acoustic signal for a given utterance (e.g. Autesserre et al., 1989; Fourcin et al., 1989). Acoustic discontinuities in the speech signal may be interpreted in terms of articulatory gestures, but such interpretations are relatively imprecise in view of the lack of a comprehensive model of acoustic/articulatory relationships.

One of the aims of the ESPRIT II - ACCOR project has been to provide a detailed description of the main articulatory and acoustic correlations in coarticulatory processes in seven European languages: Catalan, English, French, German, Irish Gaelic, Italian and Swedish. In this cross-language study we have designed a database (EUR-ACCOR which incorporates articulatory, aerodynamic and acoustic information of a variety of speech items. It has been necessary to adopt a common methodology, i.e. standardised investigative tools and normalised measurement procedures at specified locations in the speech signals. In the segmentation and labelling of the database we have adopted a non-linear approach to the annotation of articulatory, aerodynamic and acoustic events, based on the evidence provided by five different channels of information: acoustic waveform, oral and nasal volume velocity traces, linguo-palatal contact patterns, and laryngographic trace (Marchal and Nguyen, 1990).

In this paper, using EUR-ACCOR data on connected speech, we show how this truly multi-level approach may lead to a better understanding of articulatory-acoustic mapping.

THE EUR-ACCOR DATABASE

The database consists of three types of speech material: (i) VCV nonsense words; (ii) real words matching the VCV sequences; (iii) sentences constructed to illustrate the main connected speech processes in that language such as assimilations, weak forms etc. (For further details of the EUR-ACCOR database see Hardcastle and Marchal, 1990).

To record the EUR-ACCOR corpus we have designed a specialised multi-sensor workstation based on a PC which enables simultaneous digital recording of the acoustic signal and a number of additional channels for physiological and aerodynamic data. The following transducers were used in the recordings: Electropalatograph (for measuring the timing and location of tongue contacts with the palate); Pneumotachograph with Rothenberg mask (for recording volume velocity of air flow from nose and mouth) and Laryngograph (for recording details of vocal fold vibration). The acoustic signal was sampled at 20,000 Hz.

The multichannel data may be displayed and analysed using a specialised software program (EDIT) (Hardcastle et al., 1989). EDIT also provides the facility to attach labels ("annotations") to specified instants within the speech data; these can be later used in setting up a database of measurements from the speech data. Each signal can be annotated independently of the others. Discontinuities in each trace can be marked according to specific criteria (see example of annotated utterance below, fig. 2).

A total of ten repetitions of the full corpus are produced by at least five subjects in each language, making a total of approximately 12.5 Gbytes of data.

ANNOTATION AND LABELLING

Speech production is a complex process relying on coordinated gestures. It is essentially a non-linear process, such that articulatory gestures overlap, and resulting events can be associated with more than one phonetic symbol. Our annotation therefore is based on two principles: (i) the principle of non-linearity and (ii) the principle of independence of information from the different channels. Discontinuities are interpreted as indications of oncoming gestures from and towards articulatory goals. They correspond to turning points on the different traces and they are marked in the temporal domain according to precisely defined criteria. These gestures may of course have different temporal extents in the different articulatory dimensions under investigation. This further motivates the adoption of a non-linear approach to "segmentation".

ANALYSIS

To illustrate the annotation criteria we have selected the utterance "...Susan can't go..." from the complete EUR-ACCOR sentence "Fred can go, Susan can't go, and Linda is uncertain" produced at a normal conversational rate by a female Southern English speaker. The utterance was constructed because it potentially contains a number of connected speech processes such as assimilations, reductions, deletions etc. For instance, there are final alveolar consonants that can be assimilated to the place of articulation of the following stop. This reflects the production of similar utterances in natural spontaneous conversational speech.

The segmentation criteria and the annotations used are described below. Information from all levels of representation is used. For practical reasons, the following discussion on annotation assumes the phoneme as the segment for which landmarks are identified. However, the actual phonetic realisation of each phoneme is also provided.

/s/ Annotated at onset and end of friction in the higher frequencies (1.085-1.200 = 115 ms). It is difficult to annotate from the EPG data since lingual constriction is formed gradually. The peak on the airflow trace corresponds to the onset of the friction. The second (small) peak corresponds closely to the end of the friction and soon after there is a release in the constriction for the fricative in the EPG data [240] (1.200) (line 2 on oral flow trace, fig. 2; and fig. 1).

/u/ Duration from (1.211-1.322 = 111 ms), as annotated from the spectrogram. The tongue is relatively close for this vowel as shown on the EPG data. Phonetically [ʉ].

/z/ Annotated at the onset and end of the higher frequencies on the spectrogram (1.322-1.389). It is difficult to annotate from the acoustic waveform since the drop in amplitude is very gradual and there is no significant change in the shape of the pulses. The amplitude in the laryngograph trace drops between (1.328-1.380) [265-276] (lines 2, 3 on the laryngograph trace, fig. 2). At [265] (fig. 1) there is already considerable contact in the alveolar region and at [277] there is release of the maximum contact for the /z/. The laryngograph trace does not therefore capture the approach to the gesture. Annotation from the oral airflow is difficult since the first peak cannot be easily defined. There is some build-up in nasal airflow during the fricative beginning at (1.342) (line 1 in the nasal airflow trace, fig. 2).

/ə/ (1.389-1.422) annotated at onset and end of the formant structure on the spectrogram. The vowel is fully nasalised [ə̃] as evident from the nasal airflow trace which peaks at the end of the vowel (line 2 on nasal airflow trace, fig. 2).

/n/ Assimilated to the place of the following velar stop. Annotated from the spectrogram at (1.422-1.486 = 64 ms). It can be annotated from the EPG closure to the end of voicing (1.426-1.486 = 60 ms) [285-297]. Phonetically [ŋ].

/k/ Annotated from the spectrogram and the acoustic waveform (1.486-1.525 = 39 ms). The release of the oral airflow also occurs at the same point (1.525) (line 5 on the oral airflow trace, fig. 2). There is aspiration at the release (1.525-1.571) for 46 ms. The laryngograph trace shows clearly the lack of voicing during the closure. Phonetically [k^h].

/ɑ/ From (1.571-1.747 = 176 ms) as annotated from the spectrogram and the acoustic waveform. The presence of nasal flow indicates this vowel is nasalised [ɑ̟̃].

/n/ and /t/ There is no realisation of the alveolar gesture for the two phonemes which coalesce and are realised as nasal creak lasting from (1.747-1.797). Laryngealisation is indicated by the presence of irregular cycles on the acoustic waveform and the laryngograph trace. EPG data show the presence of velar closure during the period of the nasal creak [353] (1.765). The presence of a [ŋ̰] is therefore indicated by the EPG data.

/g/ Completely voiceless velar stop from end of nasal creak (1.797) to the release indicated in the spectrogram and oral airflow trace (1.825) (line 12 on the spectrogram and 7 on the oral airflow trace, fig. 2). Aspiration from (1.825-1.852). Phonetically [k^h].

/əʊ/ The diphthong and the following /ə/ from underlying /ənd/ are reduced to a long /əː/ lasting for 160 ms (1.852-2.012).

In addition to the phonological processes expected (e.g. assimilation of the alveolar nasal to the place of articulation of the following stop) examination of the spectrogram, acoustic waveform, EPG, laryngograph and oral and nasal airflow traces has shown a variety of phonetic and idiosyncratic processes. In particular, the nasal airflow trace has shown the presence of nasal airflow not only during the production of the vowel preceding a nasal but also during the fricative /z/. In addition, EPG data have indicated the presence of velar closure during the production of the nasal creak. This multi-level analysis has therefore provided further insight to the phonetic processes present.

DISCUSSION

The main observation we can draw from the multi-channel analysis of the utterance may be summarised as follows:
a) Noticeable right to left assimilatory processes involving stops (n>ŋ). The systematic comparison of the information provided by the different channels of information indicated that there was no trace left of the underlying phoneme. This has important implications for automatic alignment of a given phonemic transcription onto a speech signal. It is necessary in cases of discrepancy to take into account the phonological assimilatory processes as manifested at the articulatory and acoustic levels. A system would fail if such representations were not present. An abstract level of representation containing information such as speaker specific socio-idiolectal phonological rules as well as speaking rate, style etc. is necessary for successful recognition.
b) Nasalisation which anticipates or carries over into the actual realisation of the nasal segment. A comparison of the nasal and oral airflow is of particular interest since it indicates indirectly the movements of the velum and the control of the velopharyngeal port.
c) Complete or partial overlap of phonologically consecutive segments, e.g. /ntg/ > [ŋ̰k^h]in "can't go". The need for additional channels of information can be best illustrated in such cases of overlap where spectral information alone may not be sufficient in depicting the underlying organisation and sequence of segments.

The non-linear multi-level approach adopted in the ACCOR project is particularly well-adapted to the investigation of the simultaneous activity of the different motor subsystems and the resulting acoustic output. This activity can give a precise view of the timing relationships of the different components of the phonatory and articulatory gestures, information which is often not possible from observation of the acoustic signal alone.

REFERENCES

Autesserre, D., Pérennou, G. and Rossi, M. (1989) *Methodology for the transcription and labelling of a speech corpus*. Journal of the International Phonetic Association, 19, 1, pp. 2-15.

Fourcin, A.J., Harland, G., Barry, W., Hazan, V. (eds) (1989) *Speech input-output assessment; multi-lingual methods and standards*. Chichester: Ellis Horwood.

Hardcastle, W.J. Jones, W., Knight, C., Trudgeon, A. and Calder, G. (1989) *New developments in electropalatography: A state of the art report*. Clinical Linguistics and Phonetics, 3, pp. 1-38.

Hardcastle, W.J. and Marchal, A. (1990) *EUR-ACCOR: A multilingual articulatory and acoustic database*. Proc. 1st ICSLP, Kobe, Acoust. Soc. Jap., pp. 1293-1296.

Marchal, A. and Nguyen-Trong, N. (1990) *Nonlinearity and phonetic segmentation*. J.Acoust.Soc.Am., Suppl. 1, Vol. 87, pp.79-82.

ACKNOWLEDGEMENTS

We acknowledge financial support from The European Economic Commission DGXIIIunder the auspices of ESPRIT II Basic Research Action (ACCOR project 3279). Thanks are due to Andy Butcher , Linda Shockey and Fiona Gibbon for helpful comments on the manuscript.

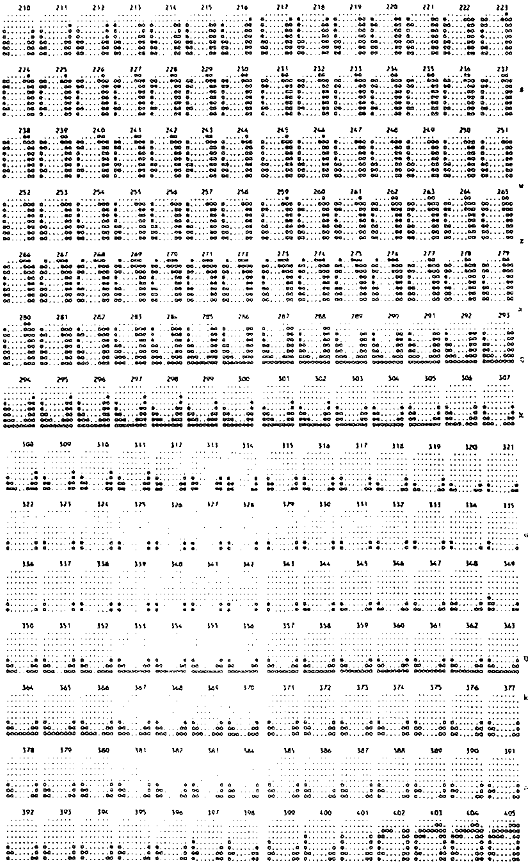

Figure 1. Full EPG printout of the utterance "Susan can't go". Interval between frames is 5 ms.

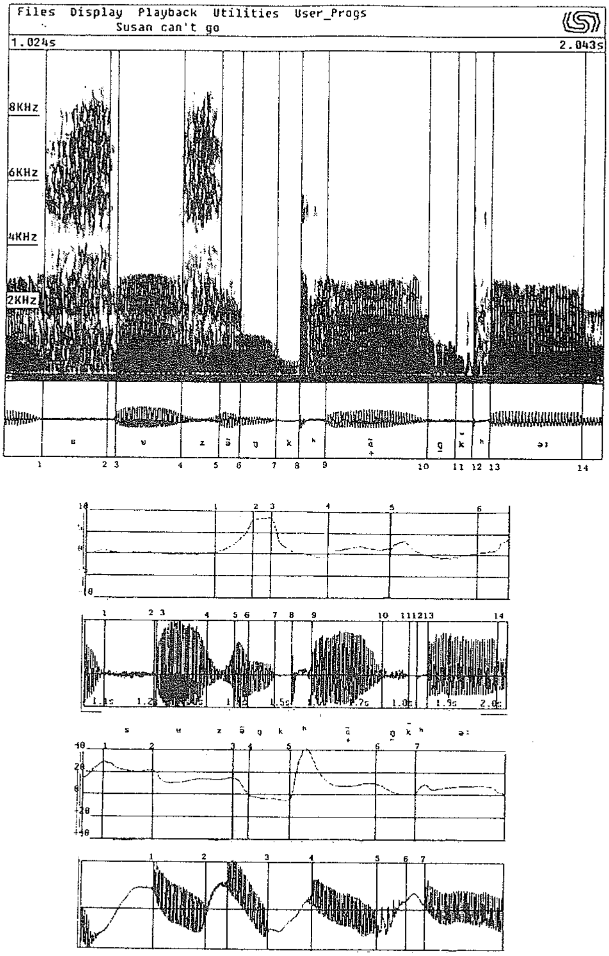

Figure 2. *(upper)* Wideband spectrogram of the utterance "Susan can't go" with segmentation lines marked numerically and phonetic transcription included (NB the attenuation of the higher frequencies is caused by the mask used for the airflow recordings). *(lower)* EDIT screen display of the utterance "Susan can't go" showing nasal airflow, acoustic waveform, oral airflow and laryngograph trace. Segmentation lines are included separately for each channel and correspond to the major discontinuities in each signal. They are numbered separately for each channel and the relevant times from the beginning of the utterance (in parentheses) and EPG frame numbers [in square brackets] are provided below.

Acoustic Waveform: 1: (1.085) [216], 2: (1.200) [240], 3: (1.211) [242], 4: (1.322) [264], 5: (1.389) [277], 6: (1.422) [284], 7: (1.486) [297], 8: (1.525) [305], 9: (1.571) [314], 10: (1.747) [349], 11: (1.797) [359], 12: (1.825) [365], 13: (1.852) [370], 14: (2.012) [402]. Nasal Airflow: 1: (1.342) [268], 2: (1.426) [285], 3: (1.472) [294], 4: (1.601) [320], 5: (1.759) [351], 6: (1.963) [392]. Oral Airflow: 1: (1.085) [216], 2: (1.200) [240], 3: (1.389) [277], 4: (1.426) [285], 5: (1.525) [305], 6: (1.731) [346], 7: (1.825) [365]. Laryngograph Signal: 1: (1.211) [242], 2: (1.328) [265], 3: (1.380) [276], 4: (1.477) [295], 5: (1.580) [316], 6: (1.740) [348], 7: (1.797) [359], 8: (1.852) [370]