X-ray films of speech

(This demo was originally prepared by Phil Hoole for a CD-ROM illustrating use of QuickTime in science produced by Dr. M. Batschkus, Klinikum Grosshadern, Munich. In the present version, the original examples are followed by links to additional xray-films from the same corpus.)

A few examples are given here of x-ray films of short sentences taken from a much larger corpus of x-ray material. These films have had a tortuous history. They were originally recorded on 35mm film in 1974 by Dr. Claude Rochette at the Departement de Radiologie de l'Hotel-Dieu de Quebec, Quebec, Canada. In the early nineties they were transferred by Drs. Kevin Munhall (Queens's University, Kingston, Canada), Eric Vatikiotis-Bateson and Yohichi Tohkura (ATR Labs, Kyoto, Japan) to analog video disk, as the original films were in danger of disintegration. In order to do this, the audio track had to be processed to remain in synch with the slower frame rate of the video disk, compared to the original films (30 vs. 50 fps.). This is why the audio track sounds somewhat strange: when the films are played at 30fps the pitch is correct but the tempo of the utterance is slower than the original.

In 2000, Phil Hoole of Munich University transferred fourteen out of the total of twenty-five films from the videodisk to computer. The digitization was performed without compression, i.e the full frame-rate and resolution were retained. (The assistance of Dr. Marc Batschkus (Multimedia Lerncenter Medizin, Klinikum Grosshadern, Munich) in transferring the films is gratefully acknowledged.) Each film consists of roughly 30 short sentences (typically about 100 frames per sentence). The image area containing the x-ray information is about 400 vertical by 500 horizontal pixels in size. Each sentence in the film has been stored as a separate file. The data is currently stored as MATLAB-compatible files, since we use MATLAB as our main processing and display environment. However, in order to make the films more generally available, especially for didactic purposes, we are now starting to convert the material to QuickTime format.

Why do x-ray films constitute an important speech resource?

X-ray filming is even today effectively the only imaging technique that allows all the most important speech articulators (jaw, tongue, lips, soft palate, larynx) to be captured in a single view at a framerate that is reasonably adequate for speech. However, in most countries it is no longer considered ethically acceptable to make cineradiographic recordings of speech with healthy subjects, unless there is some clear clinical indication for the recordings. Thus existing x-ray films form an irreplaceable source of information on speech.

What are the advantages to having the data in digital form?

Inspection of the details of speech movements often requires frame-by-frame examination of small portions of the film. Navigating in a digital film is much easier than with a normal VCR or even than with a video-disc player.

Of greater significance is the fact that once the film is in digital form, additional sources of information can be added. In the examples given here, this "added value" takes two forms:

1. A Sonagram of the utterance has been inserted into each film. A cursor moves along

the time axis of the sonagram to indicate the time point of the film frame currently visible. This is also a great help for navigation in the film, since when looking at small sections of film in slow motion it is very difficult to keep track of where one is in the utterance. In addition, understanding the relationship between speech movements and the resulting acoustics is one of the central issues in phonetics; it is thus very valuable to be able to generate such parallel displays.

2. A text track has been inserted into each film giving a transliteration of the utterance in both normal orthography and phonetic transcription (the latter using the SAMPA system; see http://www.phon.ucl.ac.uk/home/sampa/english.htm).

   Moving brackets indicate which word (in the orthography) and which sound (in the phonetic transcription) is currently visible. Again this is useful for navigation in the film (especially for those unfamiliar with sonagrams). Using the QuickTime player's "find" and "find again" functions it is possible to jump straight to specific sounds or words (maybe not so relevant in the present examples as the utterances are quite short), or simply jump through the film from sound to sound.

The ultimate aim of providing time-aligned labels of this kind is much more far-reaching, however. Once this information is available for all of the several hundred utterances in the corpus, then it would be possible to set up a searchable database with this symbolic information. Users could then search the database for specific words or sounds in specific environments (e.g all /r/ sounds, or all /r/ sounds coming after a plosive), and would receive as a result of the search a list of all relevant utterances (films) and the location in the utterances of the sounds of interest.

Example utterances

F1. Female speaker, utterance 1 (filename 180_11.mov)

"Is that biography?"

SAMPA: / I z D { t b aI Q g r @ f i /

F2. Female speaker, utterance 2 (filename 180_28.mov)

"Loraine just left"

SAMPA: / l @ r eI n dZ V s t l e f t /

M1. Male speaker, utterance 1 (filename 177_04.mov)

"Its ten below outside"

SAMPA: / I t s t E n b @ l @U aU t s aI d /

M2. Male speaker, utterance 2 (filename 177_30.mov)

"He's a lousy singer"

SAMPA: / h i: z @ l aU z i s I N @' /

Video processing: Noise suppression using an adaptive Wiener filter (mainly to remove some of the effect of the grain in the film). Gamma correction to emphasize low intensity regions. Sorensen codec.

The sonagram axes are time on the x-axis (labelled in seconds), frequency on the y-axis (labelled in Hz). Relative intensity is from black (low intensity) to white (high intensity). The bar on the right shows the mapping from grey-scale to dB.

Notes on the utterances

1. The two utterances of the female speaker give an illustration of how different the articulation of the 'same' sound can be. Contrast the r-sound in "biography" with the r-sound in "Lorraine". In the former case the /r/ is barely visible, in the latter case there is an extremely clear retroflex articulation (tongue tip curled back). This also gives a nice idea of the mobility of the tongue tip.

2. The sound sequences t-b (from "that biography"; F1) and n-b (from "ten below"; M1) are two classic cases where considerable temporal overlap in the movements of the articulators may be expected. It is worth looking at these utterances frame by frame.

In "that biography" the lip closure for /b/ follows very closely after tongue-tip closure for /t/. There is thus a period of simultaneous closure at both tongue and lips. Then the tongue closure is released, followed by release of the lip closure.

In "ten below" there is a similar pattern of movement by tongue and lips. An additional consideration is the movement of the soft palate. At the point where the lips close, the soft palate is still open (it had to be open for the /n/, of course), so in effect something like an /m/ is articulated briefly, until the soft palate raises to close the velopharyngeal port and allow air-pressure to build up for the /b/.

3. A further note on velar timing:

Nasal sounds occur in three utterances ("Lorraine", F2 ; "ten", M1; "singer", M2).

Generally, lowering of the velum occurs well before the segment that is actually labelled as nasal, usually around the beginning of the vowel preceeding the nasal, so movement of the velum in the nasal segment itself is mostly upward, which in a sense is the opposite of what one might expect.

4. Utterance M2 illustrates characteristic differences in tongue-tip configuration for alveolar consonants. For the fricatives /z, s/ the constriction is made with the tongue-blade (*laminal*), for the lateral /l/ with the tongue tip (*apical*). This again shows the mobility of the tongue tip. Note also that although the fricatives and the lateral have nominally the same place of articulation the position of the jaw is much lower for /l/.

5. Also in utterance M2, the last vowel of "singer" illustrates the r-colouring that occurs in North-American dialects of English. This is generally considered to be a very unusual vowel, since it is seen to involve two constrictions in the vocal tract, one in the palatal region (as also found in high vowels like /i/) and at the same time in the pharyngeal region (as in vowels like /a/).

Additional Examples

(The examples have sonagrams with a cursor linked to the movie, but no synchronized labels.)

L77_08 “They are a nomadic tribe”

L77_12 “McCarthy was a madman”

L77_13 “He likes Tom Sawyer the best”

L77_14 “Tom likes Greco-Roman art”