SYNCHRONIZED CINERADIOGRAPHY AND VISUAL-ACOUSTIC ANALYSIS[1]

H. M. TRUBY

In order to speak intelligently about the correlation of sound and physiological activity, I feel I must first refer briefly to the individual development of the relevant separate, analysis techniques. The past 15 years especially have been increasingly dynamic years for phoneticians, for we have had made available to us increasingly dynamic analysis instrumentation designed to help us better understand the clearly dynamic medium of communication which speech really is.

We were first impressed with sound spectrograms, and the general reaction to them was, "Now we've got speech where we want it!" But as the formal reports appeared – Potter,[2] Steinberg and French,[3] for examples, and then later the book *Visible Speech*[4] in 1947, and Joos's monograph[5] in 1948 – it had become apparent that we really didn't have speech where we wanted it, yet. The kymogram and oscillogram had posed new problems of evaluation for many years. The sound spectrogram was doing likewise, perhaps in an even more challenging fashion.

For the next unique development we can, I believe, credit the absolutely indispensable team of Cooper, Delattre, and Liberman with their program for the synthesis of speech. For myself, as Pierre Delattre assiduously tutored away, I felt, "Now speech is really hooked!" And I must admit to increasing encouragement during these last several years in Stockholm during the piecing together of the now very active laboratory of Gunnar Fant, for the synthesis phase helps immeasurably in the understanding of sound spectrograms.

The first cineradiographic and sound film available to me was the Haskins film of 1954. In Stockholm in 1957 I began to make my own films, and I soon became aware that speech was still elusive and more challenging than ever. The past few years have given me an opportunity to improve my understanding of the evaluation we are striving for – for the physical correlations possible with visual-acoustic analysis of whatever sort and cineradiography. The Doctors Subtelny have reviewed for us this

[1] As tape-recorded during the session.

[2] Potter, Ralph K., "Visible Patterns of Sound", *Bell Tel. System Monograph B-1368*, and *Science*, vol. 102, pp. 463–470 (9 Nov., 1945).

[3] Steinberg, J. C., and French, N. R., "The Portrayal of Visible Speech", *BTS Monograph B*-1415, and *JASA*, 17, 1–89 (1946).

[4] Potter, R. K., Kopp, A. G., and Green, H. C., *Visible Speech* (D. van Nostrand, N.Y., 1947).

[5] Joos, Martin, *Acoustic Phonetics, Language Monograph* No. 23 (1948).

morning certain details concerned with the techniques of radiography and have warned us fairly. I would repeat both *their* warnings and a few of my own if time allowed. Instead, I wish to direct my attention primarily to the physical operation of the correlation of sound and picture.

As X-ray film is being exposed, a multiple recording is essential. The sound should be tape-recorded and with the maximal reduction of undesirable background noise. The individual frame exposures must be synchronized with the sound track. This synchronization I have always accomplished with a direct-writing oscillograph which produces a permanent paper record. On such a record one can get the sound track on one channel and on another channel the frame marking, which gives all the ultimate correlation needed. Transfers may be made later if it is desired. And then, of course, ultimately the sound track may be subjected to various kinds of sound spectrographic and other visual-acoustic analysis. For this purpose, one can photograph a string of Sonagrams and reduce them photographically to the size wished. Before presenting the slides related to this material, I offer you two minutes of my X-ray, sound and motion film[6] which will be shown in its entirety on Thursday, along with scientific films of other members of the Congress. This will be followed by the aforementioned slides demonstrating 1) synchronization; 2) multiple, lateral laminography; and 3) multiple, frontal laminography with simultaneous lateral radiography for control. (There followed two minutes of sound cineradiography of the vocal tract, in 16-mm copy.) Now, the limitations, of course, of such a film as the foregoing are the manners in which it may be presented for analysis. Everything enters in – the size and illumination of the screen, the type of projector, the darkness of the room, the speed of projection, and so on. A really good hand projector with a variably controlled speed offers something of the dynamic opportunities that are needed for making a valid appraisal. After all, with the camera placed in the position permitting a sagittal plane view, the kind of limitations to be expected we are all familiar with. One needs a mechanically efficient hand projector allowing both forward and backward manipulation. The *backward* manipulation, especially, presents movements one is not familiar with from *any* point of view, thus calling attention to articulation which would otherwise pass unnoticed. And with *single frame* projection, one can make tracings. The variable speed control enables one to view the film to its ultimate best advantage.

The extraction of a single frame from the body of the film is also possible, and if one is fortunate enough to get at least a fairly clear delineation, this individual frame can be used for study. Without getting into details of evaluation, I wish to demonstrate the contrast between the pictured pharynxes of Figure 1 and Figure 2. Figure 1 occurs during the utterance of a word containing an American English /ɑ/, at a point in the utterance during maximal oral cavity opening and tongue flattening. Figure 2 occurs at exactly the same relative position in the utterance of a word containing /u/, same speaker, at a point in the utterance during maximal lip rounding (that is, minimal ori-

[6] Truby, H. M., "Correlation of Cineradiographic and Visual-Acoustic Analyses of Speech" (17-min., 35-mm, sound and motion, X-ray film, available in 16-mm copy) (Stockholm, 1958).

Fig. 1. Instant of maximal tongue flattening and oral cavity opening during [-ɑ-] of deliberate utterance [plɑpt]. X-ray screen photographed at 48 frames per second. Midwestern American English speaker.

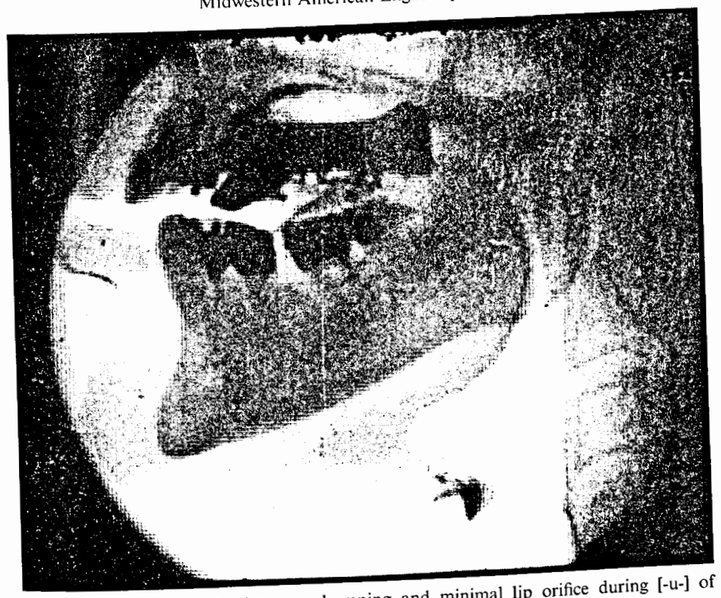

Fig. 2. Instant of maximal tongue humping and minimal lip orifice during [-u-] of deliberate utterance [plupt]. Same particulars as for Figure 1.

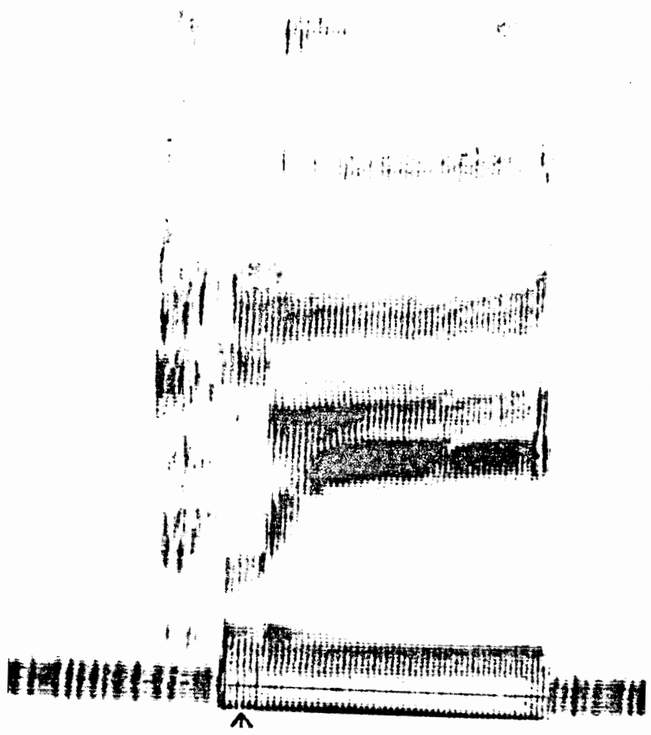

Fig. 3. Sonagram of [pli] ('plea'), Midwestern American English speaker. Arrow indicates approximate middle, along the time dimension, of so-called "voiced lateral resonance".

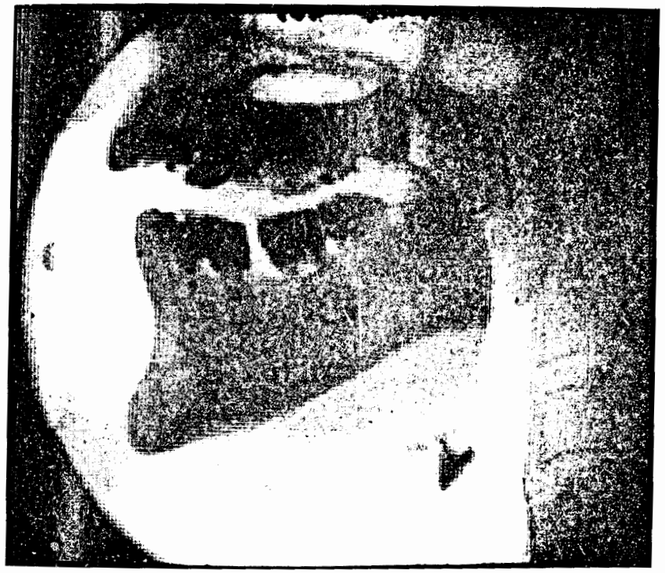

Fig. 4. Instant at approximate mid-point, time-wise, of "voiced lateral resonance" of 'chlor-' as in 'chlorine'. Same particulars as for Figure 1.

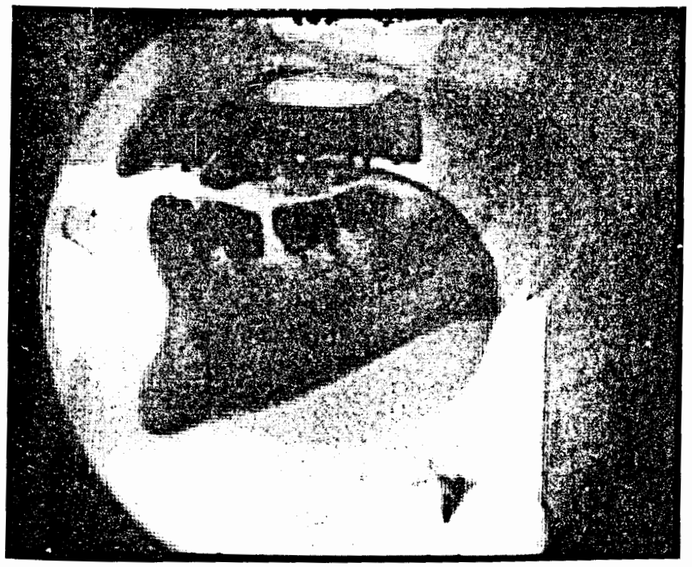

Fig. 5. Similar instant in 'blooped' as described for Figure 4.

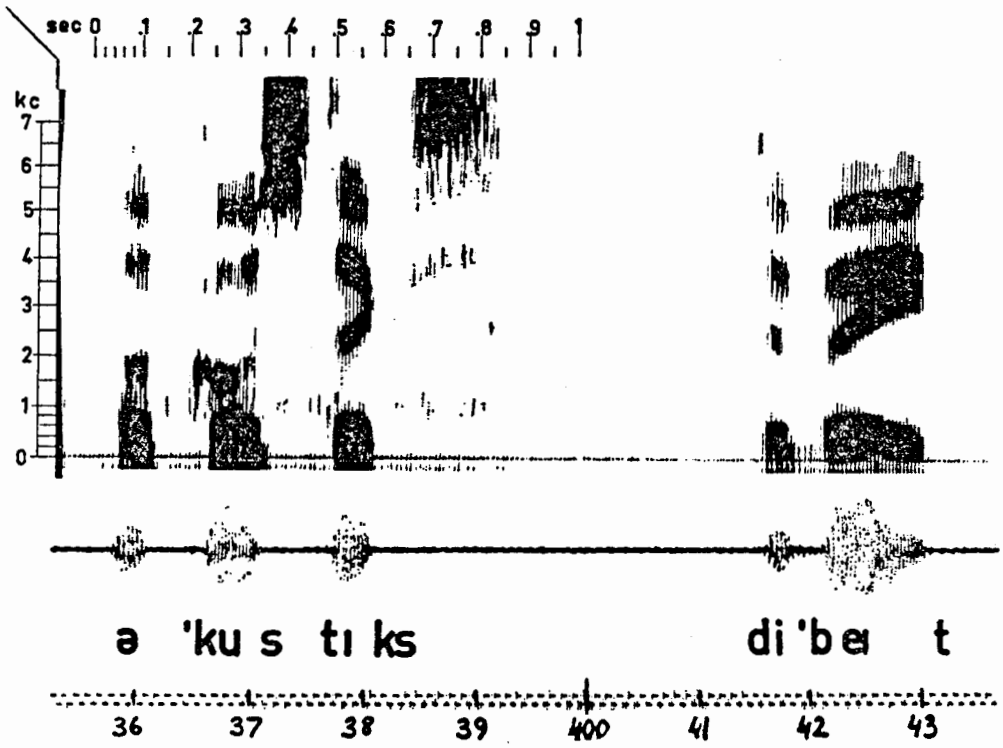

Fig. 6. Sonagram of the utterance: "acoustics, debate" correlated with synchronized Mingogram and frame-marker. The corresponding cineröntgen frames can thus be readily selected.

fice) and tongue humping.[7] We see volumetric differences on the order of an estimated 14 to 15 times increase of the /u/-pharynx over the /ɑ/-pharynx. These are startling facts in the consideration of vocal tract dimensions and indicate even further something of the complexities to be reckoned with in determining the physical properties associated with speech production.

Related to pharyngeal adjustment is the factor of physiological anticipation, reflected in, among other things, pharyngeal *pre*-adjustment. This was the point of my remark to Mrs. Dr. Subtelny this morning regarding the influence of the associated vowel, whether "in the same syllable" or not, on the preceding consonant or consonant cluster. It may be demonstrated that the physical description of any instant of a speech sound is more seriously contingent upon the precise phonetic environment – especially that which will *follow* the instant in question – than the concept ALLOPHONE ordinarily is considered to account for. For example, consider Figure 3. Here we see a Sonagram of a Midwestern American English utterance of the word 'plea'. A portion of this spectrogram is characteristically referred to as the "voiced lateral resonance". The instant I now refer to occurs somewhere in the middle, time-wise,

[7] The röntgen screen was filmed at *c.* 48 frames per second to provide these pictures.

of such resonance and is indicated, for 'plea', by the arrow, at the bottom of the picture, just opposite the striation corresponding to the fourth vocal pulse (or vocal cord vibration) which is visibly registered here. Figure 4 and Figure 5 are of two more single frames extracted from the film which was presented in part a few moments ago and are representative of two instances of the instant of lateral resonance described above. The first instance (Figure 4) occurred in an utterance (also Gen'l Am.) of the "syllable" 'chlor-', as in 'chlorine'. The second instance (Figure 5) occurred in an utterance (same speaker) of the word 'blooped'. Even though each instant is a "lateral resonance mid-point", many differences are immediately obvious: lipopening, approximation of upper to lower teeth, general oral cavity opening, position of tongue dorsum with respect to juncture of hard and soft palate, apposition of tongue root to back (dorsal) wall of meso-pharynx, general pharyngeal outline, and angle of epiglottis. In short, we see the articulators already anticipating the approaching vowel during the voiced resonance portion of the preceding lateral in each instance. It may be hoped that such revelations will elicit from investigators more attention to the influence or effect a given phone invariably has on the preceding phone. I will go so far as to say that at least 90% of all phonetic assimilation is *pro*gressive. Taking only monosyllables, for simplicity's sake, we see clear-cut radiographic and spectrographic evidence that so-called *vowel length* is a function of the final consonant(s); we see that perhaps *all* the phonic material – and indisputably "the vowel" – is nasalized by a final nasal; we see the physical nature of the initial consonant and of each "member" of an initial "consonant cluster" to be directly dependent upon the following vocalic articulation *and* – though to a lesser degree – upon the nature of the final consonantal complex. Additional aspects of progressive assimilation are functions of degree of syllabic stress, rate-of-utterance, and individual articulation habits such as degree of voicing.

And now to the actual physical synchronization of sound and picture, Figure 6. During filming, while taking a sound recording at the same time, we have the direct-writing oscillograph going, and an oscillogram is produced of the acoustic material. The words here were 'acoustics', 'debate'; the second channel of this permanent oscillogram displays the frame markings – we see indications of the 360th, 370th, 380th, and so on beyond the 430th, as well as the frames in-between; the open shutter time may be calculated from the square wave form of the continuous marking. A later correlation can be made by placing directly on the recording tape a little magnetic click. Thus can be made an accurate correlation between such visual-acoustic records as, in this case, Sonagram and Mingogram – the latter a kind of frequency-limited oscillogram. One can make as many of these oscillograms as are wished and at the original speed, the speed being constant on this apparatus. The degree of accuracy is within the time of a single vocal pulse (or vocal cord vibration cycle).

Figure 7 demonstrates what usually happens when one records against the X-ray camera noise. The Sonagram is of the word 'blooped' plus background noise. Compare it with the Sonagram of Figure 6, for the recording and filming of which latter

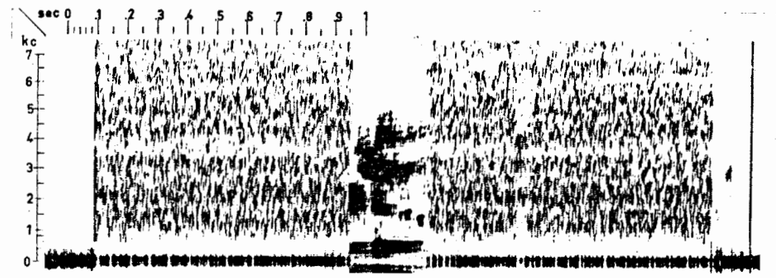

Fig. 7. Sonagram of the utterance [blupt] (female, fundamental frequency *c.* 240 cps) plus camera and apparatus noise in the "background". The final consonantal complex is clearly masked out.

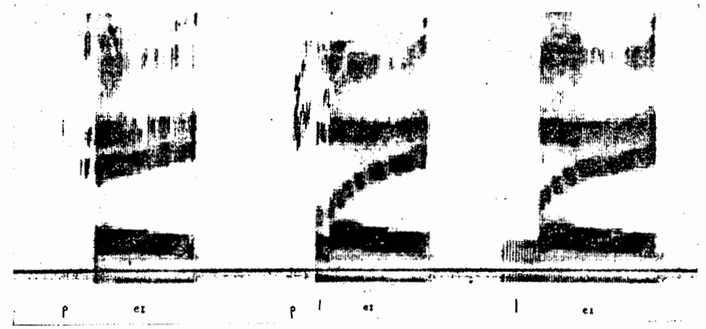

Fig. 8. Sonagram of utterance "pay, play, lay". Note differences in acoustic nature of post-explosion frications of "pay" and of "play" – the articulatory positioning for the lateral is completely accomplished *prior to* the explosion.

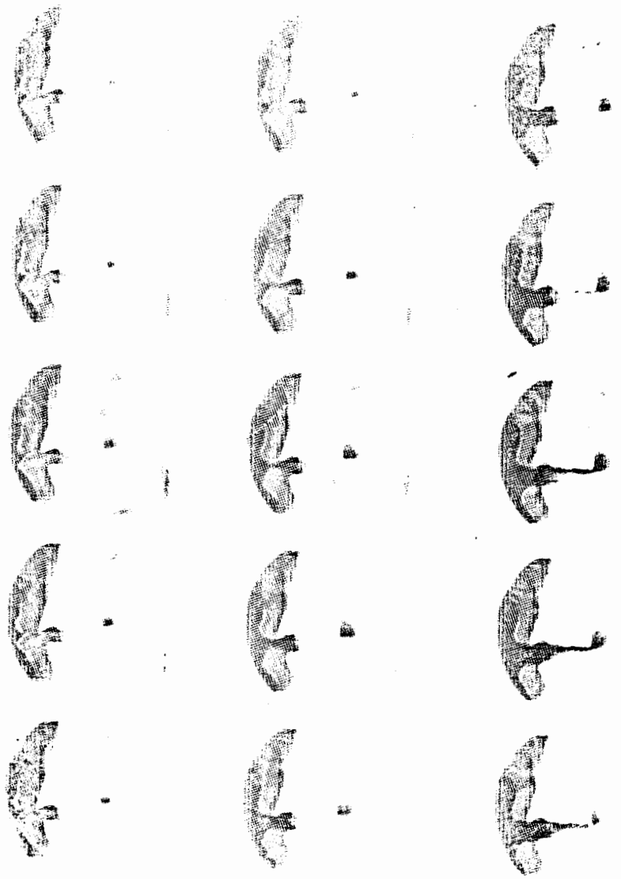

Fig. 9. Cineradiograms of certain pre-audible and audible articulations during an utterance of [plɑč], Midwestern American. Photography: 48 frames per second, as numbered.

1	6	11
2	7	12
3	8	13
4	9	14
5	10	15

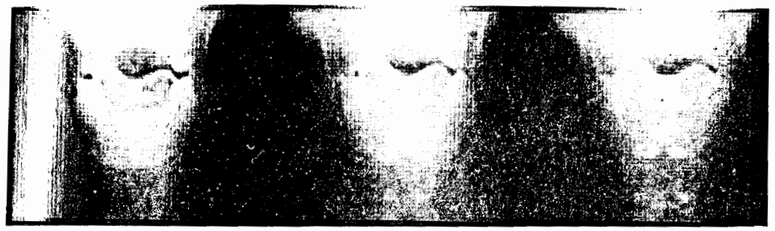

Fig. 10. Three of five multiple laminograms (simultaneous tomograms) during a sustained isolated-vowel utterance [ɑ:::] (Swedish). Note asymmetries of tongue-surfaces, roof-of-the-mouth, and oral cavity as a whole. Puzzle: select "highest point of the tongue".

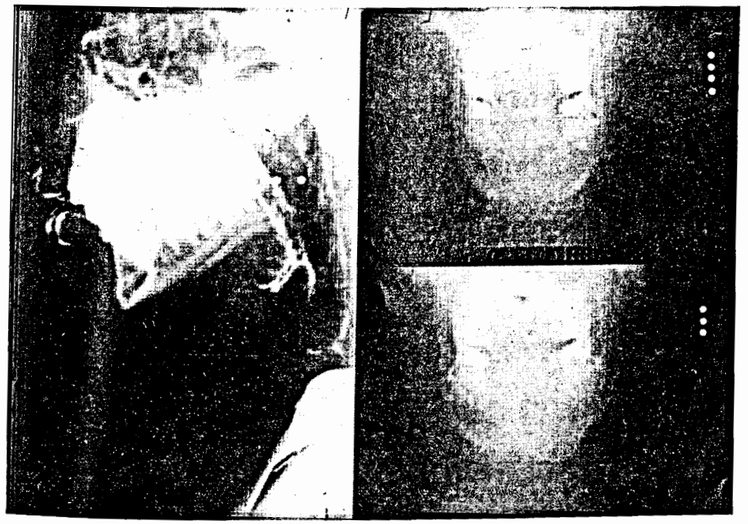

Fig. 11. Synchronized single lateral radiogram and two of five multiple frontal laminograms (frontal polytomograms) of a sustained isolated-vowel utterance [i:::] (Swedish), synchronously sound-recorded on magnetic tape.

Fig. 12. Multiple lateral laminograms (lateral polytomograms) of a sustained isolated-vowel utterance [i:::] (Swedish), synchronously sound-recorded on magnetic tape.

utterances the subject was placed in a sound-insulated chamber. The X-ray screen was inserted through the wall of the chamber, and cotton was packed around it. The use of a small-field microphone contributed to a gross reduction of 30 db of background noise on the recording as may be estimated from Figure 6.

The material of Figure 8 demonstrates what directed me to make a film in the first place. We see 'pay', 'play', 'lay'. I was interested in contributing something to the question of successivity or simultaneity in clusters. One sees, for 'lay', the voiced lateral resonance previously noted. For 'play' it is less than half so long, preceded by the so-called aspiration part. The obvious differences between the post-explosion frication in 'pay' and in 'play' led me to believe that something was happening here that would be interesting to look at cineradiographically.

The fifteen frames of Figure 9 were extracted from the film. In Frames 1 through 6 we see the lips closed for the initial occlusion of '*plotch*'. The lips have not quite come open for the plosion, in Frame 7, yet the tongue has already assumed the "most lateral position" that it will assume in the utterance. In fact, there has been no evidence of further upward movement or tensing of the tongue since Frame 3. Now, the only thing left for the tongue to do – since it's already up – is to come down. And this it does once the explosion has taken place (between Frames 7 and 8) and a certain amount of first voiceless and then voiced lateral resonance occurs. One sees the blade – but not the tip – moving downward by Frame 10, but there is still voiced lateral resonance in evidence on the correlated spectrogram. The *lingual flap*[8] portion of the pre-vocalic lateral articulation occurs, here, between *Frames 10 and 13* and corresponds on the spectrogram to the so-called "*vowel* transition" introducing [-ɑ-]. This "flapping down" of the tongue tip and blade is perhaps the primary phonetic characteristic of an initial (or pre-vocalic "in the same syllable") lateral, the resonance being of only supplementary or secondary importance to the articulation.[9] The so-called *lateral flap* comprises a *lingual flip* (upward movement) immediately followed by a *lingual flap* (downward movement). Thus, the initial plosion of [plɑč] is heard with the tongue and other articulators *already in position* for the phonemically sequential lateral. This phenomenon is discussed at length elsewhere.[10] What is true for /pl-/ is similarly true for /bl- kl- gl- pr- br- tr- dr- kr- gr- pj- bj- kj- gj- tw- dw- kw- gw-/, but *not* for /č-/ nor /ǰ-/.[11] More of the "vowel proper" articulation is revealed in Frames 14 and 15.

Figure 10 displays three of a set of five simultaneous laminograms (tomograms) taken from the front. The sustained isolated-vowel utterance was *Swedish* [ɑ:::].[12]

[8] Truby, H. M., *Acoustico-Cineradiographic Analysis Considerations, with especial reference to certain consonantal complexes*, *Acta Radiologica*, Suppl. 182 (Stockholm, 1959), p. 151–§ 3.45.
[9] *idem.*, pp. 142–§2.298, and 162–§3.48, and elsewhere.
[10] Truby, *op. cit.*, §1.632 – pp. 120–121, and Truby, H. M. and Wegelius, C. W. "A Study of Certain Speech Articulation Processes", *Proc. of III Int'l Cong. of Medical Radiophotography (Photofluorography)*, Stockholm, 1958, pp. 267–268.
[11] Truby, *op. cit.*, §4.6–p. 191, and §5.0–p. 192.
[12] This utterance was simultaneously tape-recorded.

The second of these photographic sections is *one cm further back* in the mouth than the first, and the third is similarly focus-extended from the second.[13] Note the striking tongue-surface asymmetries and try to calculate just how one should evaluate traditional "height-of-tongue" or "maximal vocal tract constriction" calculations based on laterally viewed radiograms!

Figure 11 reveals my technique for cross-checking on the frontal, multiple laminography. I simply take a single lateral projection during the planigraphic exposure. On one channel of a two-channel tape-recorder I record the utterance; on the other synchronized channel I record a buzz signal which is coordinated with the 5-second multiple laminogram exposure and a second buzz signal of a different frequency which is coordinated with the ½-second single radiogram exposure. The lines of steel spheres are one cm apart and are photographed with the subject for other obvious control measures and measurements.[14]

Figure 12 reveals that lateral laminography provides a much sharper image than frontal laminography. The lower left-hand exposure is a photographic section through some of the leftmost teeth. The upper left-hand exposure is a photographic section 1 cm further away from the camera; this focal plane creates a section apparently through the "right tongue-hump" as viewed in Figure 10 – this is partly an illusion, however, for wherever the focal plane makes a "cut" there will appear an apparent surface. The middle planar section more or less intercepts the mid-line of the tongue, though it is clear from Figure 10 that the lowest antero-posterior surface of the tongue is not necessarily coincident with the *median line* of the tongue; one sees also the somewhat lighter projection of another tongue-dorsum surface in this exposure. The upper right-hand section intercepts a contour quite different from that of the upper left-hand section. And the lower right-hand section is through some of the rightmost teeth. Close inspection of this set of exposures will reveal other interesting contour variants.

I shall conclude by having the projectionist show a bit more of the film itself, since, with these remarks as introduction, your attention has been called to two or three things to watch for. For example, the contours of the tongue revealed by laminography may be followed very clearly during the articulation. These contours or surfaces change quite obviously, and it is not difficult to imagine the contour changes that are taking place in addition to the three *obvious* ones. It may be seen too that if one speaks a list of words, the velum drops down noticeably between each word, thus opening the naso-pharynx, but no breath is taken. This is true regardless of whether the words are monosyllables or polysyllables. The velum drops down after each – a kind of an anticipatory breathing reflex, I think. If I extend this process to sentences, there will be relaxations only between the sentences, breath-grouping relevant.

[13] Truby, H. M., Törnwall, L., and Wegelius, C. W., "Bi-planar Radiography and Multiple Tomography of Vocal Tract, with Simultaneous Sound Spectrography, during Sustained Vowels", to appear in *Acta Radiologica*, 1963.
[14] Truby, Törnwall, and Wegelius, *op. cit.*

Another thing your attention is called to is the pivoting motion of the epiglottis as seen from the side quite clearly. Also, since a contrast medium was used, a certain amount of it catches in the valeculae, and one sees small white blotches between the epiglottis and the root of the tongue.

The impression I hope to have left is that current radiographical techniques, including synchronization with sound spectrography, offer much opportunity to collect physical data related to speech performance. I do not suggest that we have arrived at a stage of finished technique or ultimate methodology. The *physical* activities associated with speech and hearing remain challenging – the unambiguous *perceptual* evaluation of these performances awaits us on an even more distant horizon.

(There followed 30 seconds more of the film begun earlier.)

Wenner-Gren Research Laboratory
Norrtulls Hospital
Stockholm