

# A method for finding constrictions in high front vowels

Michel T.-T. Jackson and Richard S. McGowan

*CReSS LLC, 1 Seaborn Place, Lexington, Massachusetts 02420*

*ladmtj@ix.netcom.com*

**Abstract:** The purpose of this study was to devise a consistent and robust method for defining vocal tract constrictions in high front vowels. A procedure was devised to find the length and position of the articulatory constriction in high front vowels that is not sensitive to local fluctuations in vocal tract shape and to the constriction-defining parameters. A method based on a visual examination of plots for constriction length and position as functions of the constriction-defining parameters was found to provide stable constriction definitions.

© 2010 Acoustical Society of America

**PACS numbers:** 43.70.Bk, 43.70.Kv [AL]

**Date Received:** July 3, 2009 **Date Accepted:** October 21, 2009

## 1. Introduction

Constriction positions and lengths are important parameters in determining the acoustic output of the vocal tract for vowels and other speech sounds (Stevens, 1998; Fant, 1960). However, not many methods for automatically determining these constriction parameters from x-ray or other data have been published. As Iskarous (2005) observed, “imaging modalities such as X-ray cinefluorography, MRI, and ultrasound produce images ... but these images are difficult to quantify and compare.” Even determining the direction of tongue movement in such images (e.g., Wood, 1982; Iskarous, 2005) does not determine the actual position and size of the resulting constriction, since a small tongue blade movement toward a nearby portion of the hard palate may produce an acoustically relevant constriction, whereas even a large tongue body movement toward a relatively distant portion of the dorsal wall of the pharynx may not. This study presents a method for determining constriction parameters in high front vowels when midsagittal images are available from any of the afore-mentioned measurement techniques. The method is robust in three ways: its results are not sensitive to local speaker-specific variation in vocal tract shape, its results are not sensitive to the selection of specific values of the constriction-finding method’s own parameters, and it is applicable across a number of speakers of different languages.

Figure 1 illustrates a common kind of local speaker-specific variation in vocal tract shape, from a cineradiographic tracing of the vowel /i/ as produced by the speaker described in Perkell, 1969. The midsagittal cross-dimensions in this token are measured along gridlines constructed according to the procedure described in Jackson and McGowan, 2008. These midsagittal cross-dimensions, plotted as a function of the gridline number, are shown in Fig. 2. In this vowel token, the small undulation in the shape of the palate creates a local minimum [indicated by the arrows in Figs. 1 and 2(a)] in the vocal tract’s midsagittal cross-dimension function. Yet neither the acoustically relevant vocal tract constriction nor articulatory goal of the constriction-making gesture is likely to be limited to this local minimum alone.

It is possible to eliminate the effects of this kind of local minimum in at least two ways. One is to take the criterion for being “within the constriction” as being some multiple  $k$ , with  $k > 1$ , of the minimum midsagittal cross-dimension. Another way to eliminate the effects of a local minimum is to construct a criterion for being within the constriction based on the mean midsagittal cross-dimension  $d$  within some window around the location of the minimum. This reduces the effects of single gridlines having small cross-dimensions.

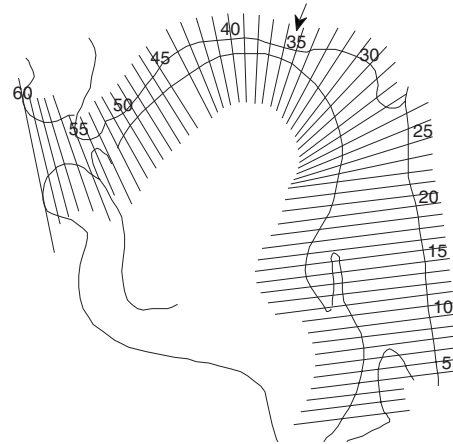


Fig. 1. Measurement gridlines for the speaker described in [Perkell, 1969](#), shown on a token of /i/. The arrow near gridline 35 shows a local minimum in the vocal tract midsagittal cross-dimensions measured along the gridlines.

This study investigates the effect determining a criterion cross-dimension  $c = k \cdot d(n)$  while varying both  $k$  and the number  $n$  of gridlines posterior to the location of the minimum cross-dimension, in order to determine a combination of  $k$  and  $n$  that is both useful in determining constriction length and position in high front vowels, and is not unduly sensitive to individual cross-dimension measurements on particular gridlines. This method will be applied to articulatory data from a variety of languages.

## 2. Procedure

### 2.1 Materials

Various images of the vocal tract during high, front vowel production from 13 different speakers were analyzed in this study. The vowels were from North American English, French, Spanish, and Mandarin Chinese. The digital image of each vowel token was enhanced and aligned using the MATLAB Image Processing Toolkit. Hard structures in and near the vocal tract, especially teeth, fillings, and the hard palate, were used as landmarks for rotation and translation of the images in the alignment procedure as described in [Jackson and McGowan, 2008](#). The images analyzed as part of a larger study are summarized in Table 1.

In North American English, tracings of cineradiographic images of three speakers' productions of the vowels /i, ɪ/ were analyzed in this study. The three speakers included the speaker from [Perkell \(1969\)](#), and two speakers from the ATR X-ray Film Database ([Munhall et al., 1994](#)) originally recorded by [Rochette \(1973, 1977\)](#).

Original tracings from [Perkell \(1969\)](#), provided courtesy of the author, were scanned and digitized at 300 dpi. These tracings were based on cineradiography at 45 frames/s. The cineradiographic recordings of running speech in the ATR X-ray Film Database ([Munhall et al. 1994](#)) have been digitized and are available on DVD. The original films were shot at 50 frames/s. Single frames from multiple tokens of the vowels /i, ɪ/ produced by the speakers L73/74 and L78/79 were selected and analyzed. Each frame was taken at or near the extreme of mandibular motion or a period of negligible tongue movement during the vowel.

In French, tracings of cineradiographic images from running speech of three speakers' productions of vowels from [Bothorel et al., 1986](#) were analyzed. The original x-rays were filmed at 50 frames/s. The tracings were digitally scanned at 300 dpi, and the vowels /i, y/ were used in this study.

In Spanish, tracings of x-ray images of three speakers' static productions of Spanish

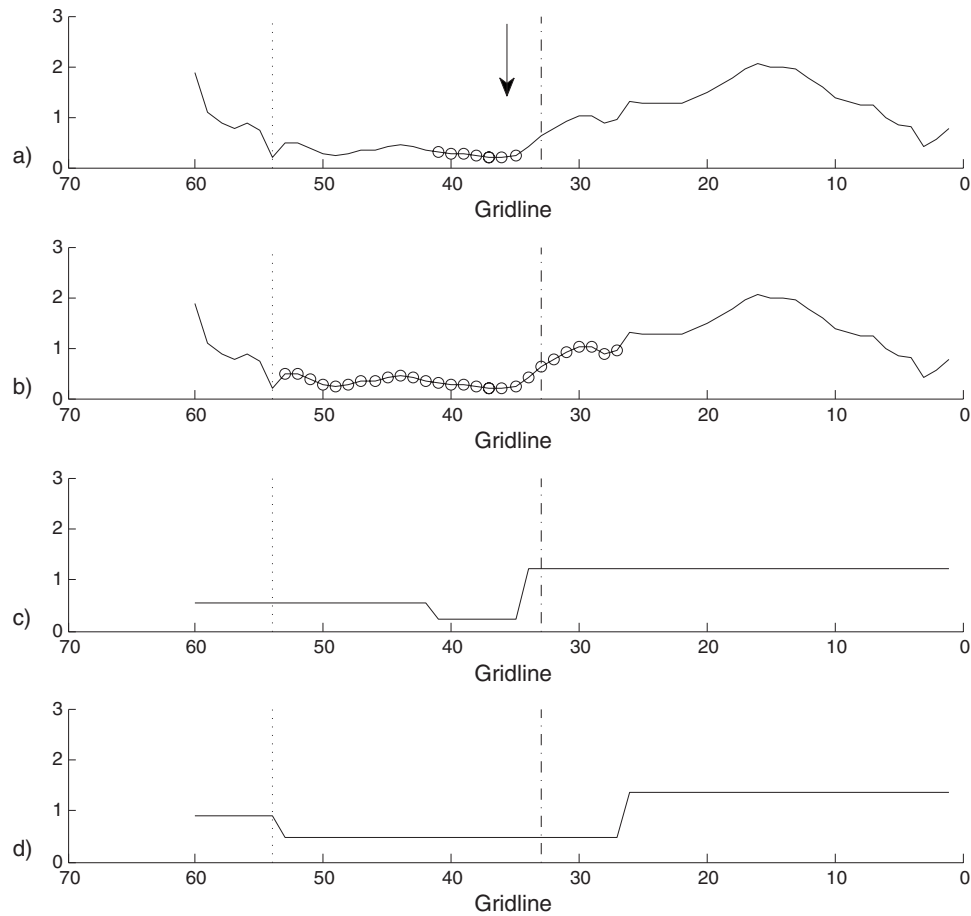


Fig. 2. The midsagittal cross-dimension function for the vowel token shown in Fig. 1, together with typical results of the parametrized constriction-finding procedure and derived three-tube models. The dashed line at gridline 33 represents the posterior boundary of the hard palate; the dotted line at gridline 53 represents the anterior boundary. Gridlines marked with “o” are in the constriction found by the procedure. The arrow near gridline 35 shows the same local minimum as Fig. 1. In (a), the o’s show the constriction found using  $k=1.5$  and  $n=2$ . In (b), the o’s show the constriction for  $k=4.0$  and  $n=8$ . (c) shows the three-tube model derived from (a); (d) shows the model from (b).

vowels from Navarro-Tomás, 1916; Parmenter and Treviño, 1932; Russell, 1929 were analyzed. Each tracing was scanned at 300 dpi, in some cases after photocopying with enlargement. The vowel /i/ was used in this study.

In Chinese, tracings of x-ray images of four speakers’ productions of Chinese vowels from Ohnesorg and Svarný, 1955; Abramson *et al.*, 1962 were analyzed. The speakers (speakers A and B from Ohnesorg and Svarný, 1955; and speakers 1 and 4 from Abramson *et al.*, 1962), spoke either Mandarin or a closely related dialect.

The tracings of x-ray images of static productions of vowels from Ohnesorg and Svarný, 1955 were scanned at 300 dpi after photocopying with enlargement. The vowels /i, y/ were used in this study. The original 16 mm film described in Abramson *et al.*, 1962, which was shot at three times the normal rate for 16 mm film (i.e.,  $3 \times 24$  frames/s = 72 frames/s) was transferred to Betamax- and VHS-format videotape in earlier work. The VHS-format videotape was then digitized and transferred to DVD. Single frames from running speech taken near the articulatory peak of the vowels /i, y/ in the digitized video were selected and analyzed in this study.

Table 1. Summary of languages, speakers, vowels, and number of tokens used to construct measurement gridlines in this study.

Language	Speaker and source	Number of tokens													
		i	ɪ	e	ɛ	æ	y	ø	œ	ʌ/ə	ɑ	ɔ	o	ʊ	u
English	Perkell	1	1	—	3	1	—	—	—	—	1	—	—	1	1
	L73/74	13	12	—	8	7	—	—	—	—	8	—	—	3	8
	L78/79	8	6	—	7	8	—	—	—	3	6	—	—	1	9
French	S1	2	—	2	2	—	2	2	1	—	2	2	2	—	2
	S2	2	—	2	2	—	2	2	1	—	2	2	2	—	2
	S3	2	—	3	2	—	2	2	1	—	2	2	2	—	2
Spanish	NT	1	—	1	1	—	—	—	—	—	1	1	1	—	1
	PT	1	—	1	1	—	—	—	—	2	1	1	—	1	1
	R	1	—	1	—	—	—	—	—	—	1	—	1	—	1
Chinese	OSA	2	—	—	—	—	1	—	—	1	3	—	—	—	1
	OSB	1	—	—	—	—	1	—	—	2	—	—	2	—	1
	A1	4	—	—	3	—	3	—	—	3	5	—	2	—	5
	A4	2	—	—	2	—	1	—	—	1	1	—	1	—	2

### 2.2 Gridlines

The measurements used in this study were taken along gridlines based on the method described in Jackson and McGowan, 2008, supplemented where necessary by gridlines spaced evenly from the alveolar ridge to the line tangent to the anterior surfaces of the upper and lower lips in the midsagittal plane. As in Jackson and McGowan, 2008, all the tokens produced by each speaker were traced to ensure that all of each speaker’s gridlines were above the highest position of the glottis. In order to make the measurements comparable from speaker to speaker, 53 gridlines were placed between the upper incisors and the highest position of the glottis.

### 2.3 Parametrized constriction-finding procedure

Since the overall goal of this study was to analyze high front vowels, which typically have a constriction in the palatal region of the vocal tract, the initial step was to manually define a region of interest for each speaker. The region of interest was taken as all the gridlines intersecting the hard palate. In Fig. 1, the posterior boundary of the hard palate is approximately gridline 33, marked with a dashed line in Fig. 2. The anterior of the hard palate is gridline 53, marked with a dotted line in Fig. 2. The gridline with the minimum midsagittal cross-dimension in the region of interest was then identified. The mean cross-dimension  $d(n)$  for the minimum plus the cross-dimensions from the  $n$  gridlines posterior to that gridline was calculated. Gridlines posterior to the minimum were used in order to avoid including the incisors and lip constriction in the window. Thus  $n=0$  is the minimum cross-dimension alone,  $n=1$  is the window with the minimum cross-dimension plus the next gridline posterior to the minimum. The number of gridlines in the window was  $(n+1)$ . The criterion distance  $c=k \cdot d(n)$  was calculated and the constriction was then defined as the set of all contiguous gridlines around the minimum gridline with cross-dimensions less than the criterion distance. The multiplier  $k$  was varied in steps of 0.25 from 1.25 to 8.0 and  $n$  was varied from 0 to 10.

### 2.4 Tube models

In order to provide some insight into the acoustic plausibility of the constriction found by the procedure, three-tube models based on the constriction found for each value of  $k$  and  $n$  were constructed. The posterior, constriction, and anterior tube diameters were taken as the means of

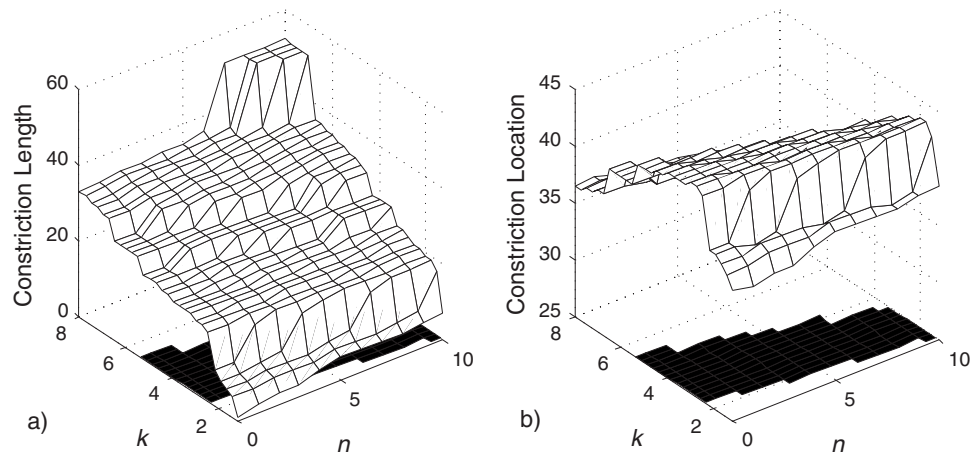


Fig. 3. The results of varying  $k$  and  $n$  over the ranges 1.25–8.0 and 0–10, respectively. (a) shows the length of the constriction (in gridlines) determined by the constriction-finding procedure. (b) shows the location of the center of the constriction determined by the procedure. The black shadow under each plot shows the range of  $(k, n)$  values that yield stable constriction length or location estimates.

the cross-dimensions of all the gridlines posterior to, within, and anterior to the constriction, respectively. The cross-dimensions were converted to areas using the function given in Wood, 1982 (p. 35)  $A = 1.93 (d^{1.5})$ , where  $A$  is the estimated cross-sectional area and  $d$  is the mean mid-sagittal cross-dimension of the relevant section of the vocal tract. For comparability with other work, the cross-sectional area of the constriction was normalized to  $0.3 \text{ cm}^2$  (see Stevens, 1998) (p. 277 ff. and Fig. 6.1) and the areas of the other sections were scaled up or down according to the same factor. The overall length of the tube model was set to 16 cm. The area function of each tube model was then used in the EASY speech synthesizer (McGowan and Wilhelms-Tricarico, 2005) in order to determine its transfer function and the corresponding formant frequencies.

### 3. Results

Figure 2 shows typical results of the parameterized constriction-finding procedure. Figure 2(a) shows the results for small values of  $k$  and  $n$ , and Fig. 2(b) shows the results for larger values. It can be seen that as  $k$  and  $n$  increase, the constriction identified by this procedure grows. Figure 2(c) shows the tube model corresponding to the constriction as identified in Fig. 2(a); Fig. 2(d) shows the tube model corresponding to Fig. 2(b). The EASY synthesizer calculated formant frequencies  $F_1 = 338 \text{ Hz}$ ,  $F_2 = 1819 \text{ Hz}$ , and  $F_3 = 2140 \text{ Hz}$  for the tube configuration in Fig. 2(c); and 393, 1912, and 2643 Hz for Fig. 2(d).

Figure 3 summarizes the results of varying  $k$  and  $n$  through the entire range. Figure 3(a) shows the length of the constriction (in gridlines) determined by the parameterized constriction-finding procedure. For values of roughly  $2.5 < k < 4.5$ , the plot shows a “shelf” on which the constriction length is not very sensitive to changes in  $n$ . As  $k$  increases beyond about 5, depending on  $n$ , the constriction found by this procedure expands to eventually include the entire vocal tract, a length of 53 gridlines. A black shadow under Fig. 3(a) shows the range of  $(k, n)$  under the shelf. Smaller values of  $k$  and  $n$  can be rejected on acoustic grounds (see below).

Figure 3(b) shows the location of the center of the constriction. As  $k$  increases, the constriction location again reaches a stable shelf around gridlines 40–45, but when  $k$  is greater than about 5, the constriction moves to a more posterior location than would be expected. On the other hand, the location of the center of the constriction determined by this procedure is relatively insensitive to the value of  $n$ . In Fig. 3(b), the shadow shows the range of parameter values under the shelf. Again, smaller  $k$  and  $n$  can be rejected on acoustic grounds (see below).

Figure 4 shows the results of synthesis of each candidate tube model resulting from the  $(k, n)$  combinations. The values of  $F_1$ ,  $F_2$ , and  $F_3$  identified by the EASY synthesizer for each

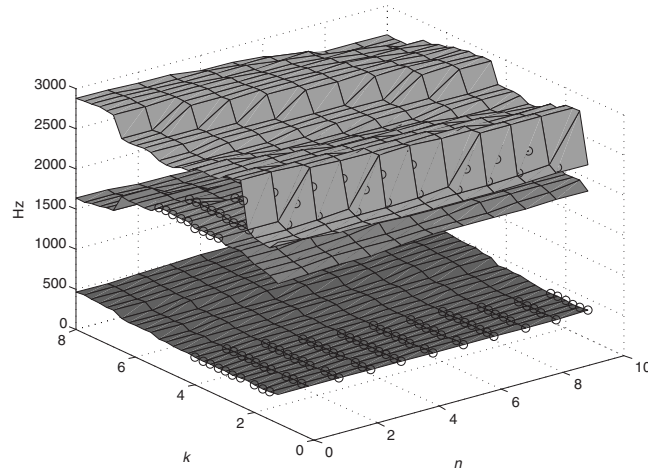


Fig. 4. Results of EASY synthesis of each three-tube model derived from the constriction-finding procedure.  $F_1$ ,  $F_2$ , and  $F_3$  are plotted as a function of  $k$  and  $n$ . On the  $F_1$  and  $F_2$  surfaces, points highlighted with an o emphasize the regions around the minimum  $F_1$  and maximum  $F_2$  values. The regions are values of  $(k, n)$  where the value of the formant was within 2.5% of the minimum  $F_1$  or maximum  $F_2$ .

three-tube model (i.e., each combination of  $k$  and  $n$ ) are plotted.  $F_1$  varies from 330 to 620 Hz,  $F_2$ , varies from 1560 to 2000 Hz, and  $F_3$ , varies from 2136 to 2910 Hz. The figure is annotated with symbols showing the range of  $k$  and  $n$  values which produced tube models that satisfied various acoustic constraints appropriate to /i/. On the  $F_1$  and  $F_2$  surfaces, points highlighted with an “o” are  $(k, n)$  combinations where the value of  $F_1$  was within 2.5% of the minimum  $F_1$ , or maximum  $F_2$ , respectively. The percentage was chosen solely for illustrative purposes to emphasize the regions enclosing the ranges of points with the minimum  $F_1$  and maximum  $F_2$ .

In Fig. 4, as in Fig. 3, it can be seen that the results are not very sensitive to  $n$ , but vary more in the  $k$  direction. Generally, realistic values of  $F_1$  in the “basin” around the minimum value of  $F_1$  are limited to  $k \leq 3$ . Similarly, the  $F_3$  surface shows that values of  $k < 2.5$  give unrealistically low values of  $F_3$ . Utilizing these constraints on the value of  $k$ , it turns out that the plateau of high  $F_2$  values crosses the  $k=3$  line at  $n > 2$ . However, since the acoustic-to-articulatory mapping itself is well-known to be indeterminate (Atal *et al.*, 1978), it should not be surprising that constraints on the values of  $F_1$ ,  $F_2$ , and  $F_3$  are not sufficient to uniquely determine the parameters of the procedure for finding the constriction in a given  $x$ -ray profile.

The articulatory analysis was repeated for every high front vowel token. The results were then combined by counting the number of times each combination of  $k$  and  $n$  fell within the “plateau” or “shelf” for the specific token, as in Fig. 3. The results, weighted equally by speaker, are shown in Fig. 5. Since vowel tokens from 13 speakers were analyzed in this study, combinations of  $k$  and  $n$  in Fig. 5 that are larger, and approach  $N=13$ , are better for more speakers. In both Figs. 5(a) and 5(b), it can be seen again that the results are not very sensitive to the  $n$ , as long as  $n \geq 3$ . In the range  $n \geq 3$ , values of  $k$  that include the most tokens appear to fall in the range  $2.5 < k < 3.5$ .

#### 4. Discussion

The method for defining vocal tract constrictions in this study is based on finding a contiguous set of gridlines around the gridline with the minimum cross-dimension. It has been shown that the method gives consistent results over certain ranges of two parameters: the number of cross-dimensions averaged with the minimum cross-dimension and the multiplier. It is not dependent on specific values chosen. Because the threshold cross-dimension used to define the constriction is based on a window of gridlines, the method is not sensitive to local variation in vocal tract shape. It has been tested across a number of speakers of different languages.

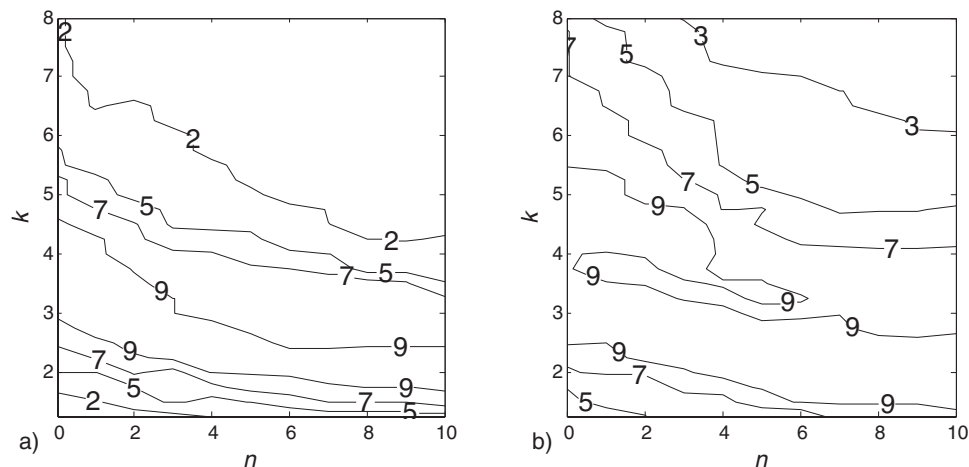


Fig. 5. The number of times each combination of  $k$  and  $n$  yielded appropriate constriction estimates. (a) shows the results for constriction length. (b) shows the results for constriction location. Results from each speaker weighted equally, so that the maximum value (not actually attained) could have been  $N=13$ .

Pragmatically, it seems like this method is easiest to understand with a small window size. Values of  $n=3$  and  $k=3.0$  seem suitable. These values also fall within the range of parameters that yield acoustically plausible three-tube models of the vowels in question.

### Acknowledgment

This work was supported by Grant No. NIDCD-001247 to CReSS LLC.

### References and links

- Abramson, A. S., Martin, S., Schlaeger, R., and Zeichner, D. (1962). *Mandarin Chinese X-Ray Film in Slow Motion with Stretched Sound* (Columbia University, Columbia-Presbyterian Medical Center and Haskins Laboratories, New York).
- Atal, B. A., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.* **63**, 1535–1555.
- Bothorel, A., Simon, P., Wioland, F., and Zerling, J.-P. (1986). *Cinéradiographie des voyelles et consonnes du Français (Cineradiography of French vowels and consonants)* (Institut de Phonétique de Strasbourg, Strasbourg).
- Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague, Netherlands).
- Iskarous, K. (2005). "Patterns of tongue movement," *J. Phonetics* **33**, 363–381.
- Jackson, M. T.-T., and McGowan, R. S. (2008). "Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels: Statistical considerations," *J. Acoust. Soc. Am.* **123**, 336–346.
- McGowan, R. S., and Wilhelms-Tricarico, R. (2005). "An educational articulatory synthesizer, EASY," *J. Acoust. Soc. Am.* **117**, 2543.
- Munhall, K. G., Vatikiotis-Bateson, E., and Tohkura, Y. (1994). *X-Ray Film Database for Speech Research (ATR Human Information Processing Research Laboratories, Kyoto, Japan)*.
- Navarro-Tomás, T. (1916). "Siete vocales Españolas (Six Spanish vowels)," *Revista de Filología Española (Review of Spanish Philology)* **3**, 51–62.
- Ohnesorg, K. and Svarný, O. (1955). *Études Expérimentales des Articulations Chinoises. (Experimental Studies on Chinese Articulations.)* (Czech Academy, Prague), Vol. **65**, Issue No. 5.
- Parmenter, C. E. and Treviño, E. (1932). "An X-ray study of Spanish vowels," *Hispania* **15**, 483–496.
- Perkell, J. S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge (MIT, Cambridge, MA).
- Rochette, C. (1973). *Les Groupes de Consonnes en Français (Consonant Groups in French)* (Laval University Press, Quebec).
- Rochette, C. (1977). "Radiologie et phonétique (Radiology and phonetics)," *Vie Médicale au Canada Français (Medical Life in French Canada)* **6**, 55–67.
- Russel, G. O. (1929). "The mechanism of speech," *J. Acoust. Soc. Am.* **1**, 83–109.
- Stevens, K. N. (1998). *Acoustic Phonetics* (MIT, Cambridge, MA).
- Wood, S. (1982). "X-ray and model studies of vowel articulation," *University of Lund Phonetics Laboratory Working Papers* **23**, 1–49.