ON THE ANALYSIS AND SYNTHESIS OF VOWELS AND FRICATIVES

JÁNOS MÁRTONY
C. CEDERLUND, J. LILJENCRANTS, B. LINDBLOM

The information transmitted in speech communication passes through several successive stages in the transmitter and in the receiver. The first stage or level in the transmitter is a psychological (or linguistic) level. The second stage is physiological. Here the information is coded neurologically, i.e., in the form of nervous activity which controls articulation here including both phonation and supraglottal events. The movements of the vocal organs produce acoustic signals. This is the physical level.

Granted a knowledge of the code it should be possible to describe the signal unambiguously at any specific stage of the communication chain. Our knowledge and technical facilities do not yet permit us to do this equally well for all levels.

At the articulatory level we can describe the movements and positions of the vocal organs as a function of time by means of, e.g., cineradiography, and we also possess the knowledge sufficient for interpreting fairly well the importance of these movements. The phonatory mechanism and the acoustic correlates of its function have been investigated only to a smaller extent.

At the acoustic level the instrumental techniques have developed more intensely and spectral data can be extracted and be described mathematically with a small number of parameters. The optimal specification and the smallest number of parameters are obtained if parameters are selected so as to closely parallel the production of speech.

This type of description does not only fulfill requirements for simplicity but also for reproducibility. The success of speech synthesis witnesses to the truth of this statement.

In vowel production there is excitation of the vocal tract at one end, i.e. at the vocal folds. Having made a distinction level we can in a similar way distinguish between source spectrum – corresponding to phonation – and the transfer function of the vocal tract corresponding to articulatory modification.

Figure 1 illustrates this division of vowel spectrum into source and transfer function. At the left-hand side of this figure are shown the larynx source characteristics in terms of an oscillogram and a spectrum. In the middle the vocal tract transfer function and its individual constituents are shown, at the right we see the oscillogram and the resulting spectrum of the complete vowel.

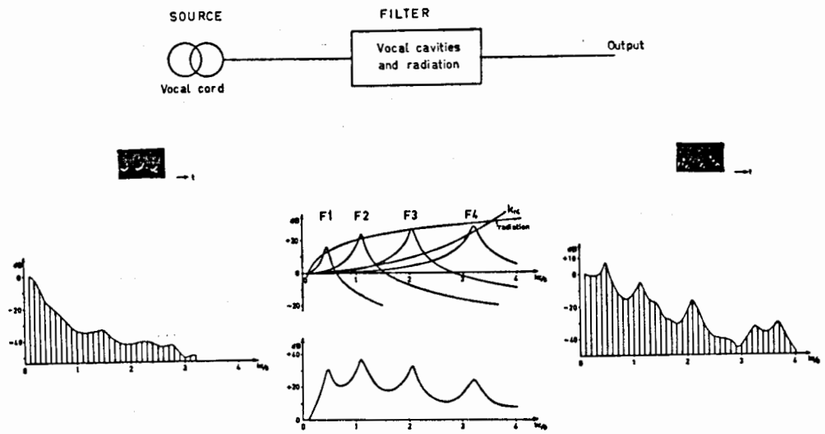

Fig. 1. Simplified block diagram of vowel production. Oscillogram and spectrum of the larynx source. The transfer function; spectrum of elementary parts F1 F2 F3 F4. Higher-pole correction k_{r_4} and radiation and the resultant sum of all constituent parts. Oscillogram and spectrum of the output, i.e. the complete sound.

The transfer function consists of a series of resonances. In the range up to 4 kc/s there are normally for male speakers 4 resonances (in mathematic terminology poles) and thus corresponding formants in the sound spectrum. To give an exact mathematical description of the transfer function the frequency position of these poles should be known. Moreover, a correction factor (higher pole correction k_{r_4}) for the neglect of higher poles than the fourth must be taken into account. The transfer function also includes the frequency dependency of the radiation from the lips of the speaker. The higher-pole correction and the position of F_4 vary inversely with the length of the speaker's vocal tract. The variations with respect to the articulation of a vowel are fairly predictable from the frequencies of the lower formants. This means that data on the first 3 formants and their bandwidths enable us to reconstruct the vocal tract transfer function. After the source spectrum has been added we obtain the spectrum of the complete sound. It should be recognized, however, that source spectra show individual and contextual variations, related to voice quality, stress, and phonemic context.

Figure 2 exemplifies larynx spectra pertaining to different degrees of a speaker's voice effort. The spectra are harmonic. In the case of normal voice effort the source spectrum falls at a mean rate of approximately 12 dB/oct. Superimposed irregularities occur frequently in the source spectrum. As a general rule an increase in voice effort is accompanied by a relative increase in high frequency energy. A spectral minimum at 800 c/s is also fairly common. An exact mathematical description of the source spectrum is of course possible to carry out but leads to rather complicated

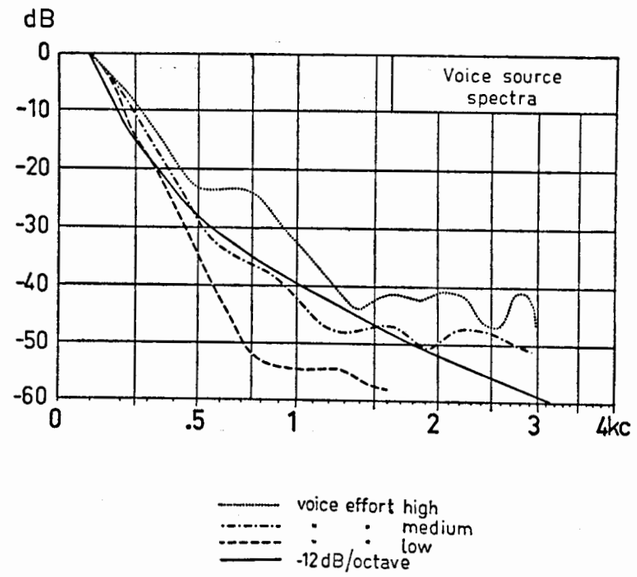

Fig. 2. Spectrum of voice source at various voice efforts.

expressions. The phonetic quality of a vowel is, however, fairly independent of the normal variations in the source spectrum.

A decomposition in terms of the source spectrum and the filtering effect of the vocal organs is also useful for the study of fricatives. The source is not located at the end of the vocal tract and has a flat spectrum (white noise). The fact that the source is not at the extreme end means that there are not only poles (resonances) but also zeros (anti-resonances visible as deep dips in the spectrum). Together with any pole originating from the posterior part of the vocal tract, i.e. from cavities situated behind the constriction, there is also a zero. If they are close enough to nearly neutralize each other we speak of a bound pole-zero pair. In the case of a larger separation they are termed free poles and zeros. The cavities in front of the constriction contribute to the transfer function only in the form of free poles.

There is a certain simplification in assuming that there is no coupling between cavities in front of and behind the constriction where the source is located. This simplification is valid, though, for high degrees of constriction. Moreover, we regard the source as a point. Model experiments indicate, however, that the source may have a finite spatial distribution and that several sources may be present simultaneously. This makes the exact mathematical treatment in terms of poles and zeros difficult in some instances. However, simple pole-zero approximations of fricative

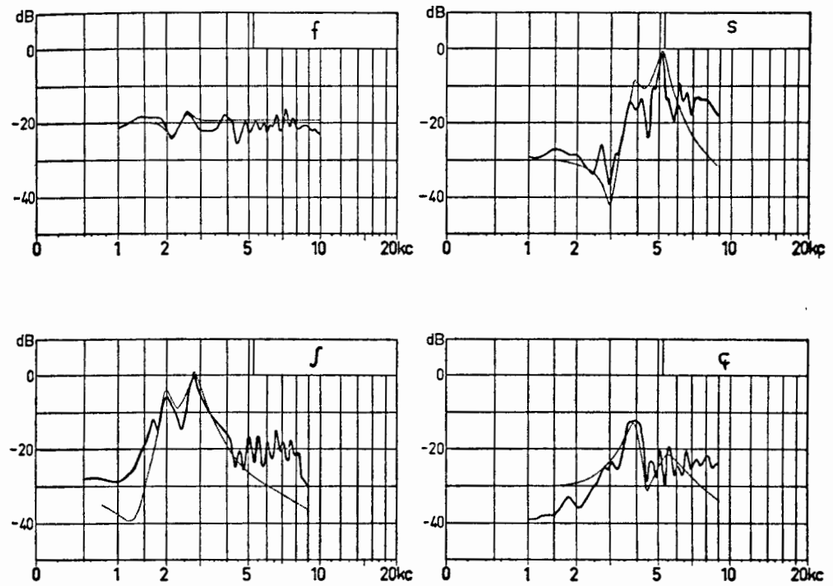

Fig. 3. Spectrum of Swedish voiceless fricatives (analysis filter B = 125 c/s τ = 80 msec). Solid line measured spectrum; dotted line 2 pole-1 zero approximation.

spectra are very useful since the deviations from the ideal method generally have no or very small perceptual significance.

In Swedish there are 4 voiceless fricatives

[f] labiodental	[s] dental
[ʃ] prepalatal (generally retroflex)	[ɕ[palatal

Figure 3 shows the spectra of these sounds.

The analyses were made of VCV-words in which V stood for [a]. All the utterances were taken from the same speaker. Spectral sections were made using 125 c/s analysis filter and 80 msec time constant.

The spectra of Figure 3 can be described in terms of the poles and zeros indicated by Table 1. These matched spectra agree with the measured spectra within $\pm$ 5 dB which could be regarded as a good approximation.

Various experiments done at M.I.T. and at the Speech Transmission Laboratory of the R.I.T. show, however, that the accuracy of a specification of fricatives based on 4–5 poles and 3–4 zeros is larger than necessary – and that an approximation limiting the parameters to 2 poles and 1 zero (dotted lines of Figure 3) is sufficient

TABLE I

	[f]			[s]	
	f c/s	B c/s		f c/s	B c/s
zero	2150	250	z	2400	250
pole	2450	250	p	2800	250
zero	3400	300	z	3100	250
pole	3700	300	p	3750	250
zero	4750	450	z	3950	250
pole	5400	450	p	4200	250
			z	4350	250
			p	4700	250
			p	5200	250
	[ʃ]			[ç]	
	f c/s	B c/s		f c/s	B c/s
z	1250	250	z	1450	250
p	1650	250	p	1800	250
z	1750	250	z	2100	400
p	2000	250	p	2800	400
z	2300	250	z	3300	300
p	2500	250	p	3500	300
p	2900	250	p	4000	300
			z	4500	600
			p	5600	700

for practical synthesis work. A preliminary test of synthetic versions of the above-mentioned VCV-words in which the fricative segments were synthesized with good accuracy and with certain approximations (cf. above) shows that listeners cannot make consistent discriminations between the two types of synthesis.

Figure 4 shows the frequency patterns of poles and zeros of the approximated fricative spectra. The following observations can be made. [f] often exhibits a fairly flat spectrum in the frequency range below 5 kc/s i.e. this region contains only bound pole-zero pairs, and at 10–15 kc/s there is a free pole. In the present [f] the free pole is not so evident. The 2.5 kc pole-zero pair is bound and appears to be influenced by coarticulation. As a general rule it is the flat spectrum that is characteristic of [f] sounds.

In [s] there is generally a zero of 3 kc/s and 2 poles at 4 and 6 kc/s resp. These poles may often be found higher up near 5 and 7 kc/s resp.

[ʃ] displays a configuration of poles and zeros similar to that of [s]. The entire spectrum is, however, shifted downwards by an octave.

In the palatal fricative [ç] the pole-zero pattern is different. There is a zero at 4.5 kc/s between the two poles.

We have attempted to show that, in the case of vowels and fricatives, the pole-zero

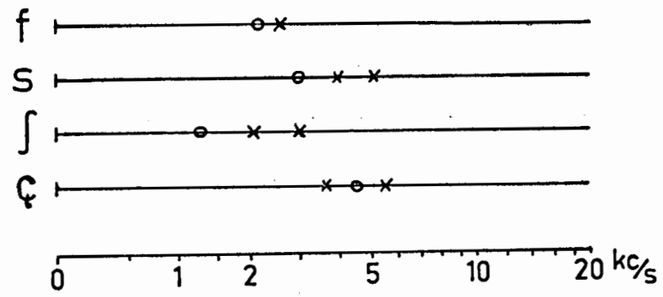

Fig. 4. Pole-zero diagram of the fricatives of Fig. 3 pertaining to the approximate spectra.

specification of spectra yield results that are not only simple but also highly reproducible. Moreover, we have reported some pole-zero combinations characteristic of the 4 Swedish voiceless fricatives. It appears that 2 poles and 1 zero provide high-quality approximations.

Speech Transmission Laboratory
Royal Institute of Technology
Stockholm

LITERATURE

(1) Fant, G., *Acoustic Theory of Speech Production* (Mouton & Co., 's-Gravenhage, 1960), 323 pp.
(2) Fant, G., "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", *Ericsson Technics*, Vol. 15, No. 1 (1959), pp. 3–108.
(3) Heinz, J. M., and Stevens, K. N., "On the Properties of Voiceless Fricative Consonants", *J. Acoust. Soc. Am.*, Vol. 33 (1961), pp. 589–596.