RECENT DEVELOPMENTS IN THE RESEARCH OF THE STRUCTURE OF VOWEL SYSTEMS

L.J. Bonder
Hogeschool van Amsterdam, dept. Logopedie/Akoepedie, Amsterdam

L.F.M. ten Bosch
Institute of Phonetic Sciences, University of Amsterdam, The Netherlands

ABSTRACT

The structure of vowel systems is discussed from the articulatory and acoustic point of view. Also, the relation between acoustic and perceptual properties of vowels is briefly dealt with. It is shown that the positions of monophthongs in a vowel system can be modelled in several ways by means of 'ease of articulation' and 'sufficient contrast'. The important models as found in the literature (e.g. vowel dispersion, Quantal Theory) are discussed in more detail.

1. INTRODUCTION

One of the most important means of communication between human beings, speech, is actually produced by accurate regulation of the subglottal air pressure and manipulation of the shape of the vocal tract. Consonant and vowel systems in languages are organized such that specific linguistic as well as phonetic (articulatory and perceptual) constraints are met: consonants and vowels serve as linguistic units but must be pronounceable and sufficiently contrastive at the same time. In this presentation, we present an outline of the present theories aiming at a structural description of vowel systems in relation to articulatory models. It is split up into three sections (**2**) the relation between articulation and perception, (**3**) the boundary of the vowel space, and (**4**) the internal structure of vowel systems.

2. RELATION ARTICULATION - PERCEPTION

For a phonetic description of vowel systems, one must look for articulatory, acoustic and perceptual differences between vowels that possibly underlie the phonological oppositions between them. At the phonetic side, we have to consider (2.1) the articulation-to-acoustics mapping, and (2.2) perceptual aspects of vowel sounds.

2.1 Articulation-to-acoustics

The computation of acoustic output from the vocal tract shape and the inverse problem, the computation of the shape from the output, constitute one of the main topics in speech production research. Most of the speech production models are based upon the source-filter theory. It has been demonstrated that speech generated by means of such models is hardly discriminable from natural speech ([11]).

The problem, how to relate vocal tract shape and acoustic output can be tackled in different ways: (1) in terms of electric LC-circuits; historically, this has been the usual paradigm originating from transmission engineering. (2) in terms of the n-tube representation of the tract. An n-tube is a tube with n segments with length l_i and area A_i. An example of this approach is given by nomograms [10], and the 'fibre'-concept [1, 5, 7]. It is also applied in the Quantal Theory ([22, 23]). (3) in terms of articulatory-based tract models. This approach, followed by Mermelstein, Maeda, Lindblom, Sundberg a.o., is characterized by the choice of a small subset of 'higher level' articulatory parameters that rule the 'lower level' tube parameters. (4) in terms of eigenfunctions of the Webster horn equation. This approach has been dealt with by [12], and, in a more loose mathematical way, in the distinctive regions theory, in [19].

These four approaches are in fact equivalent, as can be seen by considering the mathematics involved. However, all models have different starting points. Mathematically, the filter behaviour is fully described by:

$$S(z) \cdot A(z) = O(z)$$

where $S(z)$ and $O(z)$ denote the z-transforms of the input and output signal, respectively, and $A(z)$ the inverse filter. Ideally, this filter includes radiation, wall losses, etc. Such a point of view, however, only represents the technical aspects of the articulation-to-acoustics relation. It does not yield a quick insight into the relation between articulators and acoustic features.

Determination of the tract shape, while the acoustic output is given, is equivalent to decomposing $O(z)$ into the factors S and A. It is because of a priori assumptions and knowledge about these factors that such a decomposition makes sense. Thermal and viscous losses, wall vibrations and radiation effects are important when modelling the articulation-to-acoustics relation in detail. On the other hand, main effects can clearly be demonstrated by relatively simple tract models ([4, 5, 9, 10]).

An accurate calculation of the acoustic output of the tract is numerically quite involved but essentially straightforward. The *inverse problem*, however, is much harder to solve. In fact, up to now, the problem has been dealt with succesfully in the case of (sustained) vowels. In the dynamic case, the dynamic gesture is often reduced to a sequence of static articulatory positions.

It is well known that the inverse problem has no unique solution [1]. The acoustics-to-articulation relation is 'one-to-many'. The solution space, i.e. all vocal tracts producing the same acoustic output, is said to define a 'fibre' in the articulatory space. This fibre concept is well known in the general mathematical theory of mappings. In order to specify one unique exemplar from the fibre, additional constraints have to be defined. These constraints may be on the acoustic side (e.g. additional constraints on bandwidths), or on the articulatory side (e.g. minimality of an articulatory effort function).

2.2 Acoustics and perception.

Apart from the question how tube shape is related to acoustic output, the correspondence between acoustic output and perceptual features has also drawn much attention in the past two decades. This latter relation is prominent in the discussion on the structure of vowel systems. Ultimately, the structure of vowel systems is determined by linguistic (perceptual) oppositions between phonemes, and their (allophonic) realisations are bounded by physiological constraints. It has been shown that the structure of actual vowel systems is based upon the principle of perceptual contrast ([15, 16]; also [6, 9]). Several approaches have been suggested in the field of speech production. In [19] a model is proposed in which the first three eigenvectors of the Webster horn equation play a role in the determination of distinctive regions along the vocal tract. Due to the accentuation on symmetry along the tract and the criticism it not being able to describe dynamic details in some CV transients (/da/, /di/, du/) ([4]), this theory of distinctive regions still seems susceptible for some improvement. Strictly, the derivation of the results of [19] is valid in a neighbourhood of the neutral tract only; one cannot draw conclusions for more deviant formant positions.

A theory which dynamically combines articulatory gestures and acoustic output is put forward in the Quantal Theory (QT, [22, 23]). In its pure form, this theory states that the articulatory positions of which the acoustic output (in a way) is less sensitive to articulatory deviations are preferable to other positions (articulatory plateaus). The Quantal Theory predicts, in the case of vowels, the vowels that are likely

be a member of a vowel system. The presuppositions of the Quantal Theory, however, still lead to discussion and have been questioned by many authors (cf. Journal of Phonetics, vol. 17), whereas the results are not convincing (cf. e.g. [8, 14]). It is generally believed, however, that the speech signal inherits 'quantal' phonetic properties as a consequence of non-linearities of the articulation-to-acoustics mapping and probably, the categorical perception of speech sounds. If quantality exists, it is probably a result of close approximations of formant frequencies ([2, 14, 21, 23]). In [2], the importance is stressed of the difference between F_3 and F_2 (instead of the classical F_2) as a classifier between front and back vowels. Approximations of formant frequency values are called 'focal points'; there is a relation between these focal points and the notion of 'plateaus' in the Quantal Theory. The cardinal vowels correspond to focal points with respect to the F_1 and F_2 (in case of /u/) and F_3 and F_2 in case of /i/ and /a/.

It may be clear that for a proper theory of the structure of vowel systems, based upon articulatory, acoustic and perceptual features of vowels, relations must be established between very different spaces each with their own metric, any mapping between them introducing non-linearities. We must relate the phonological observations of vowel systems, with the linguistic notion of opposition as a primary tool, with the psycho-physical properties of the human hearing system, with its spectral integration and masking behaviour. In this long sequence, we have to simplify the mappings we encounter on the way in order to be able to handle all relationships.

3. BOUNDARY OF THE VOWEL SPACE

In [17], the notion of "possible speech sound" is elaborated. Phonetically, the set of possible speech sounds is a subset of the total sound-producing potential of the vocal tract. From a phonological point of view, however, 'possibility' is a function of segmental features that are relatable to articulatory and perceptual attributes. Phonologically, the boundary of the vowel space is determined by the features [low], [back, round] and [front, spread], corresponding to the cardinal vowels /a/, /u/ and /i/, respectively. Along the dimensions [height], [backness] and [rounding], all other vowels take an intermediate position. Since the vowel coordinates on these dimensions are not uncorrelated ([back] is positively correlated with [round], [low] with [central], etc.), the dimension of the set of vowels in the phonological 3D space is somewhere between 2 and 3, rather than 3. Phonological analyses, however, are not capable to explain the actual boundary in the phonetic vowel space.

By using Maeda's statistical analyses of articulatory positions ([18]) it has been shown ([20]) that the boundary of the vowel triangle can adequately be simulated by putting specific lower and upper bounds to the tube segment areas. In [5] and [9], this phenomenon is studied by using the n-tube as articulatory model. These studies confirm that articulatory models using the 4-tube are capable of showing relevant details of the mapping articulation-to-acoustics ([4, 5, 9, 10]). In particular, the boundary of the vowel space in the 2D formant space can be described in terms of articulatory constraints. By examining the contour lines of opening degree of lossless 4-tubes, it is shown that the lowness of vowels is determined by this parameter [6]. The inverse problem can always be solved uniquely – in the lossless as well as in the lossy case – by constraining the tube shape ([9]) by means of an articulatory effort function, similar to the one applied in [1].

4. INTERNAL STRUCTURE OF VOWEL SYSTEMS.

Apart from the question how tube shape is related to acoustic output, the correspondence between articulation, acoustic output and perception has drawn much attention in the past two decades, particularly in the discussions on the structure of vowel systems. This structure may be considered to be determined by the articulatory possibilities and constraints on the one hand, and the perceptual demands on the other. One of the rules which vowel systems seem to obey is

the principle of perceptual contrast.

In [15], such a rule was implemented in a computational model for the prediction of vowel systems demanding *maximal* perceptual contrast. The maximal contrast had been established by the minimization of

$$\sum_{i=1}^{N}\sum_{j=1}^{i-1}\frac{1}{D_{ij}^2} \qquad (1)$$

where D_{ij} is the Euclidean distance between any two vowels i and j in the perceptual space; N denotes the number of vowels in the system.

The results of the computation show among others an abundance of high (central) vowels for higher values of N. The L&L-model has been further elaborated in [3], considering a perceptual distance measure based on the difference between the so-called auditory spectra of any two vowels i and j. In their paper, B&L adopt two ways of interpreting the concept of auditory spectrum: (a) as loudness density pattern, and (b) as auditory filter output. The metric of the perceptual measure is no longer Euclidean as in (1) but generalized to the L_p-metric.

The computations have again been carried out under the constraint of maximal perceptual contrast. The modifications appeared to lead to a reduction of the number of high vowels for either interpretation of the auditory spectrum.

The B&L-approach has also been applied under the constraint of *sufficient* contrast in order to compute the best 50 N-vowel systems for some values of N. The frequency of occurrence has been computed for each vowel independently and it turned out that there is a tendency towards more contrast as N increases.

Both the L&L and the B&L-model do not take into account the articulation of vowels. The implementation of a simple articulatory model like the n-tube model (where $n \leq 4$) already accounts for the most prominent results obtained for much larger values of n (cf. [6, 7]). In [6] a method is proposed for the prediction of modal N-vowel systems (i.e. the collection of most occurrent systems for each N). The method is based on the assumption that modal N-vowel systems (denoted as MVS_N) are hierarchically ordered:

$$\mathrm{MVS}_{N+1} = \mathrm{MVS}_N \cup NV$$

where NV is the set of one vowel that is the most contrastive with all vowels in MVS_N. The boundary of the discretized 2D vowel space is determined by matching to normalized vowel data ([16]). 'Repelling forces' have been defined in the articulatory space A as well as in the vowel space F:

$$d_A(v_1, v_2) > A/\sqrt{N} \qquad (2)$$

$$d_F(v_1, v_2) > B/\sqrt{N} \qquad (3)$$

where A and B denote constants, and N the number of vowels. If $\mathrm{MVS}_3 = \{a, i, u\}$, it turns out that the logarithmic vowel space (especially up to $N = 6$) gives the best results (cf. fig. 1). It seems that the implementation of articulatory constraints is mainly important for the definition of the vowel space *boundary*, although the constraints may be used to model non-modal systems that contain more 'interior' vowels.

In [7], a vowel system model is proposed that is based on maximal acoustic contrast together with a minimal articulatory effort criterion. The vowel system quality parameter Q is defined as

$$Q = D_A^2 + S \cdot (D_F - 1)^2$$

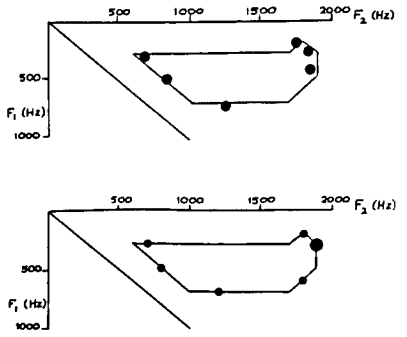

Figure 1: Vowel system prediction; (a) shows the (normalized) modal 6-vowel system; (b) shows the predicted system.

where D_A is the total articulatory system effort, D_F is the total perceptual system discrimination, and S a slack variable as used in optimization problems (S being a large positive number). D_F is computed by means of the intervowel confusion probability between two vowels

$$p(v_1, v_2) = \exp(-\alpha \cdot d_F(v_1, v_2))$$

where α is a scaling factor. Also this model, which is more fundamentally based upon probability arguments, is able to explain the main properties of vowel systems.

Vowel system models may be further elaborated by implementing submodels that describe in a more sophisticated manner the non-uniformity of the articulation-to-acoustics relation and the perceptual contrast. Studies in the 3D formant space, performed in [9] and [21], show the great dependency of the resulting model systems on variations in parameters controlling the perceptual distances between vowels.

The paper of [13] suggest a refinement of the metric used for the measurement of the perceptual contrast between nearby vowels. One of their results show that the best metric for nearby vowels is the 2D Euclidean metric after bark transformation of F_1 and F_2. Another important goal is the refinement of the overall articulation-to-acoustics relation. The Quantal Theory (QT; [22, 23]) gives us some qualitative insight into the non-uniformity of this mapping. Although the name QT is rather misleading as the relation is *continuous*, it shows quite clear its message that the acoustic change per 'unit' of articulatory change is nor uniform over the entire formant space, nor isotropic in each point of the space. As the anisotropy is greater towards the /u/ and /i/ edge of the vowel space, QT might help to increase the goodness-of-fit of vowel system models with respect to high vowels.

Acknowledgements

This study has been supported by the Dutch Organization for the Advancement of Pure Scientific Research NWO (project 300-161-030).

5 REFERENCES

[1] ATAL B.S., 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. p. 1535 - 1555.

[2] BADIN, P., PERRIER, P., BOË, L.-J., and ABRY, C. (1990). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. J. Acoust. Soc. Am. 87. p. 1290-1300.

[3] BLADON, R.A.W. & LINDBLOM, B. (1982). Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. 69. pp. 1414-1422.

[4] BOË, L.J., and PERRIER, P. (1990). Comments on 'Distinctive regions and modes: A new theory of speech production', by M. Mrayati, R. Carré, and B. Guérin. Speech Communication 9. p. 217 - 230.

[5] BONDER, L.J. (1983). The n-tube formula and some of its consequences. Acustica 52, p. 216 - 226.

[6] BONDER, L.J. (1986). A prediction method for modal n-vowel systems. Procs. Inst. Phon. Scs. Amsterdam, vol. 10. p. 73-90.

[7] TEN BOSCH, L.F.M., BONDER, L.J., and POLS, L.C.W. (1987). Static and dynamic structure of vowel systems. Procs. 11th Intern. Congress Phon. Scs., vol. 1. p. 235-238.

[8] TEN BOSCH, L.F.M. and POLS, L.C.W. (1989). On the necessity of quantal assumptions. Questions to the Quantal Theory. Journal of Phonetics 17. p. 63 - 70.

[9] TEN BOSCH, L.F.M. (1991). On the structure of vowel systems. An extended dispersion model. PhD-thesis (in preparation). University of Amsterdam, The Netherlands.

[10] FANT, G. (1960). Acoustic Theory of Speech Production. Mouton & Co., 's-Gravenhage.

[11] HOLMES, J.N. (1973). Influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. IEEE Trans. Audio-Electroacoust. AU-21. pp. 298-305.

[12] KARAL, F.C. (1953). The analogous acoustical impedance for discontinuities and constrictions of circular cross section. J. Acoust. Soc. Am. 25. p. 327 - 334.

[13] KEWLEY-PORT, D., and ATAL, B. (1989). Perceptual differences between vowels located in a limited phonetic space. J. Acoust. Soc. Am. 85. p. 1726-1740.

[14] LADEFOGED, P. and LINDAU, M. (1988). Modeling articulatory-acoustic relations. UCLA Working Papers 70. p. 32 - 40.

[15] LILJENCRANTS, J. and LINDBLOM, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. Language 48, p. 839 - 862.

[16] LINDBLOM, B. (1986). Phonetic universals in vowel systems. In: Experimental Phonology (J. Ohala and J. Jaeger, eds.). Academic Press, Orlando, Florida. p. 13 - 44.

[17] LINDBLOM, B. (1990). On the notion of "possible speech sound". Journal of Phonetics 18. p. 135 - 152.

[18] MAEDA, S. (1979). An articulatory model of the tongue based on an statistical analysis. J. Acoust. Soc. Am. Suppl. 1. Vol. 65, S 22.

[19] MRAYATI, M., CARRÉ, R., and GUÉRIN, B. (1988). Distinctive regions and modes: A new theory of speech production. Speech Communication 7. p. 257 - 286.

[20] PERRIER, P., BOË, L.J., MAJID, R., and GUÉRIN, B. (1985). Modélisation articulatoire du conduit vocal: exploration et exploitation. In: Proceedings of the 14^{emes} Journées d'Etudes sur la Parole (Groupement des Acousticiens de Langue Française, Paris). p. 55 - 58.

[21] SCHWARTZ, J.L., BOË, L.J., PERRIER, P., GUÉRIN, and ESCUDIER, P. (1989). Perceptual contrast and stability in vowel systems: a 3-D simulation study. Proceedings of the Eurospeech Conference, Paris. p. 63-66.

[22] STEVENS, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In: Human communication: A unified view (E. David and P. Denes, eds.). McGraw-Hill, New York. p. 51 - 66.

[23] STEVENS, K.N. (1989). On the quantal nature of speech. Journal of Phonetics 17. p. 3 - 45.