OXFORD PSYCHOLOGY SERIES • 23

Vowel Perception and Production

B. S. ROSNER
and
J. B. PICKERING

OXFORD SCIENCE PUBLICATIONS

Vowel Perception and Production

B. S. Rosner
Department of Experimental Psychology
University of Oxford

and

J. B. Pickering
IBM UK Scientific Centre
Winchester

OXFORD PSYCHOLOGY SERIES
No. 23

OXFORD NEW YORK TOKYO
OXFORD UNIVERSITY PRESS
1994

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland Madrid
and associated companies in Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press Inc., New York

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication Data
Rosner, B. S.
Vowel perception and production/B. S. Rosner and J. B. Pickering.
(Oxford psychology series; no. 23)
Includes bibliographical references and index.
1. Vowels. 2. Phonetics. 3. Speech perception.
4. Psycholinguistics. I. Pickering, J. B. II. Title.
III. Series.
P233.R67 1994 414—dc20 93-47078
ISBN 0 19 852138 3

Typeset by Cotswold Typesetting Ltd, Gloucester

Printed in Great Britain on acid-free paper by
Bookcraft Ltd., Midsomer Norton, Avon

Preface

The last 50 years have witnessed a rapid growth in our understanding of the articulation and the acoustics of vowels. Concomitantly, numerous investigators have produced a variety of experimental results on the perception of vowels. We originally intended to bring together this literature, along with some new findings of our own, in the present book, *Vowel perception and production*. We particularly wanted to emphasize the perception of vowels. Contemporary work on, and theories of, speech perception may strike the casual reader as concentrating on consonant perception, perhaps even on the perception of stop consonants. We hoped to correct any such imbalance through a systematic and critical consideration of the rich findings on vowel perception. Our coverage of the literature extends into the first 3 months of 1993.

As this book slowly took shape, the outlines of an auditory theory of vowel perception emerged. We have organized our presentation around that theory. The theory aims at being computational. It tries to account for vowel identification in the face of acoustic variation due to differences between speakers, coarticulatory processes, and changes in speaking rate and stress. The theory is incomplete at the algorithmic level. We have tried to indicate the steps needed to flesh it out. Whether or not the effort succeeds, we hope to have pointed the way towards new experimental and computational studies of vowel perception.

We gratefully acknowledge the use of facilities at the IBM UK Scientific Centre, Winchester, and at the Phonetics Laboratory of the University of Oxford. Some of our own results presented here are based on our recordings published in the *Oxford acoustic phonetic database*, which was partially supported by a contract from the IBM UK Scientific Centre. Helena K. K. Stoward helped in the initial assembly of bibliographic material. She also conducted Experiment 5.2 reported in Chapter 5. Nancy C. Waugh read an earlier version of the manuscript. Her sharp eye caught numerous infelicities, large and small, that we corrected in accordance with her comments.

Above all, we thank our wives, Nancy and Sarah, for their unfailing support and encouragement—and for their occasional bemused tolerance—during the writing of this book. Without their help, it would have been a far harder task.

Oxford B. S. R.
April 1994 J. B. P.

Contents

1 Introduction

1.1 Two aspects of vowel perception

Vowel perception entails two processes. One is the categorization of different vowels. An English listener hears two different vowels when, for example, a speaker pronounces the words 'bead' and 'bid'. The different meanings of those words depend on the English phonemic distinction between the vowels /i/ and /ɪ/.[1] We designate the question of how listeners manage to identify the different vowels of a language as the *vowel categorization problem.*

The second process in vowel perception is the identification of the same vowel under different circumstances. When a given vowel is produced in the same context and with the same stress by different speakers, such as a man and a woman, the outputs at their lips show quite dramatic differences. Furthermore, English listeners hear identical vowels in a single speaker's utterances of the words 'bid' and 'lid'. Rhyming, amongst other phenomena, depends on this fact. But when a speaker pronounces a given vowel in such different consonantal environments, the acoustic output at her lips during the production of the vowel usually varies with context. Nevertheless, listeners perceive the same vowel in different consonantal environments. Furthermore, producing a given vowel in the same environment but with different degrees of stress causes substantial variation in acoustic output. Other causes of variation in the acoustic output from a single speaker are speaking rate, the fundamental frequency of the voice, presence or absence of whispering, and even inherent variation within one speaker producing the same vowel in the same environment and with the same stress on different occasions. We group these latter four factors together under the heading of *momentary speaker characteristics.*

All these acoustic variations in different productions of a given vowel raise a second question. How do listeners hear the same vowel despite the vicissitudes of the physical stimulus itself? In psychological terms, this issue is a constancy problem. We therefore label the second question the *vowel constancy problem.* In part, this problem springs from *between-speaker* differences. Consonantal environment, stress, and momentary speaker characteristics are *within-speaker* sources of variation that also create the vowel constancy problem. The standard way of attacking constancy problems is to search for invariants,

---

[1] We use square brackets to designate a particular realization of a speech sound by a human speaker. Slashes are used to designate phonemes (classes of speech sounds that contrast within a language). We also use slashes in symbolizing synthetic stimuli.

either in the physical stimulus itself or in some psychological transform of that stimulus. This approach has been used in various attempts to solve the problem of vowel constancy. An obvious issue springs up at once. Can a single type of representation, physical or psychological, cope with the effects of the major sources of variation in vowels: speaker identity; consonantal environment; stress; and momentary speaker characteristics?

Understanding vowel perception therefore requires answers to the twin problems of vowel categorization and vowel constancy. Before dealing directly with these problems, we first discuss two preliminary, interrelated topics. One concerns the several techniques that have been used to represent vowels as physical stimuli and as perceptual events. The other concerns basic aspects of vowel production, since articulation determines the physical properties of vowels. In Chapter 2 we will examine the relationship between vowel production and vowel acoustics in some detail. Interpreting the various physical representations of vowels, however, requires some knowledge of vowel production.

Accordingly, in the next part (Section 1.2) of this chapter, we briefly review the source–filter theory of vowel production, which will assume a key role in Chapter 2. In Section 1.3 we cover different, currently used forms of stimulus representation. One particular representation, the $F2/F1$ plane, leads to a brief consideration of articulatory differences between vowels. In Section 1.4, we initially characterize the vowel categorization and vowel constancy problems in terms of the $F2/F1$ plane. We next discuss (Section 1.5) different auditory transforms that have been proposed to relate pitch to frequency. These transforms have been used in perceptual representations of vowels. Finally, the chapter closes with a summary of the plan of the book (Section 1.6).

1.2 Source–filter theory

As physical stimuli, vowels arise from a speaker's articulatory movements. In order to understand the physical properties of those stimuli, a brief overview of vowel production is necessary. For the present this is best done in terms of the source–filter theory of speech production.

Source–filter theory (Fant 1960) treats the articulatory system in acoustic terms. The theory is sketched in Fig. 1.1 for vowel production. The vocal apparatus is divided into two sections, the larynx and the supralaryngeal vocal tract. For most speech sounds, air is pushed out from the lungs (pulmonic egressive).[2] For vowel production the momentary actions of the larynx modify

---

[2] It is possible, however, to produce speech sounds as air is taken into the lungs (pulmonic ingressive). In some languages this is an integral part of articulation. We will consider only speech sounds where pulmonary action is egressive.

egressive airflow, providing a driving source at the top of the larynx to the supralaryngeal tract.

The source can assume various forms. In one form, the vocal folds execute quasiperiodic vibrations, going repeatedly from near approximation or complete closure to full opening and back again to near approximation or closure. The result is a quasiperiodic variation in volume velocity at the glottis. The figure illustrates this form of glottal activity which results in voiced vowels. In another condition the vocal folds are partially open and relaxed. The volume velocity at the glottis becomes aperiodic and noisy. Under this condition vowels are whispered. Other types of laryngeal action combine quasiperiodic vibrations of the vocal folds with quite incomplete glottal closure, as in breathy voice. For a given type of laryngeal activity the glottal waveform can vary between individuals (Monsen and Engebretson 1977; Price 1989).

Each type of laryngeal action produces a particular time-domain pressure waveform at the top of the larynx. Source–filter theory assumes that the source waveform due to a given mode of vocal-fold action adopts a constant *shape*, no matter what vowel is being produced. The supralaryngeal vocal tract is then treated as a filter acting on the driving waveform. This action yields a new time-domain pressure waveform at the lips. The filter characteristics of the supralaryngeal tract are assumed to be independent of the form of laryngeal activity. (This is not strictly correct, as we shall see later.) Altering the positions and shapes of the articulators produces different vowels. These operations manipulate the shape and therefore the filter characteristics of the supralaryngeal tract. In turn, changes occur in the time-domain pressure waveform at the lips. The shape of this waveform varies across different vowels, given any particular glottal waveform. Figure 1.1 illustrates one such

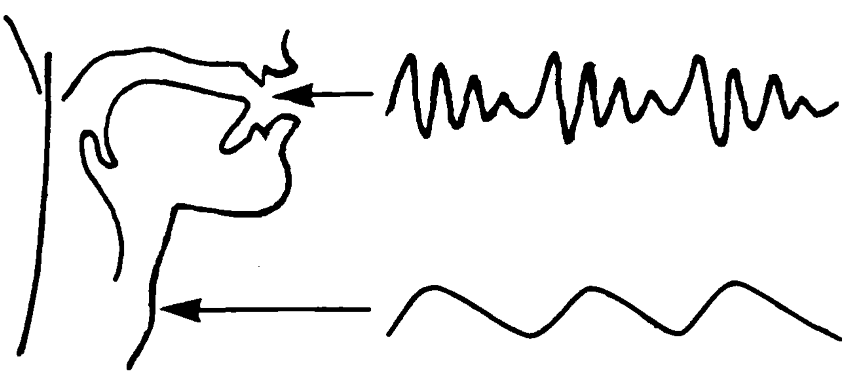

Fig. 1.1 Waveform at the glottis and at the lips during production of a vowel. After Borden and Harris (1984); reproduced by kind permission of Williams & Wilkins Co, Baltimore.

output waveform for a voiced vowel. It has the same fundamental frequency as the glottal waveform but assumes a different shape in the time domain.

In frequency-domain terms the filtering effects of the supralaryngeal tract depend on its resonant characteristics. Changes in the position of the articulators—tongue, lips, and jaw—alter the vocal-tract resonances or formants. Each formant has a centre frequency and a bandwidth, both of which vary with the positions and shapes of the articulators. Changes in the resonant properties of the supralaryngeal tract cause variations in vowel quality. Human vocal tracts express some five to seven different formants during vowel production.

Figure 1.2 recasts source–filter theory in frequency-domain terms. Figure 1.2(a) shows the spectrum of a periodic glottal waveform during the production of a vocalized vowel. The spectrum is idealized as a line spectrum. Figure 1.2(b) represents the spectral envelope of the filter characteristics created by some particular configuration of the supralaryngeal tract. The spectral envelope that results from the filtering of the source by the vocal tract and the spectral envelope of the output radiated beyond the lips appear in Figs. 1.2(c) and 1.2(d), respectively. The former is the envelope of the product of the source spectrum and the filter characteristics. The final output envelope displays the effects of frequency radiation at the lips, which tilts up the envelope of Fig. 1.2(c). An actual spectrum after vocal-tract filtering and lip radiation would show the harmonics of the glottal source. Finally, Fig. 1.2(e) shows the radiated time-domain waveform corresponding to Fig. 1.2(d).

The physical basis for vowel quality can be conceived either in time-domain or frequency-domain terms. Compact treatment of differences in time-domain waveforms, however, is no simple task. It is far easier to treat differences in vowel quality in the frequency domain. As a first step, then, vowel categorization will be related to the properties of the output spectrum. Exactly how to characterize those properties will be a matter for later chapters.

Frequency-domain treatment of source–filter theory provides some immediate, simple insights into one aspect of the vowel constancy problem. Variations in the frequency of vocal-fold vibrations shift the harmonics of the glottal spectrum but leave intact the overall shape of this spectrum.[3] The filter characteristics of a given configuration of the supralaryngeal tract supposedly remain constant. Therefore, under normal speaking conditions, the envelope of the output spectrum retains its shape, so that perceived vowel quality remains constant under changes in the perceived pitch of the speaker's voice. A change from a voiced to a whispered vowel supposedly affects the source spectrum with no alteration in the vocal-tract formants. In fact, some small

[3] Upward and downward movements of the larynx that typically occur in speech also contribute to the control of the fundamental frequency of the source spectrum. These movements, however, also modify the shape of the pharyngeal cavity, thereby affecting vowel spectra. Section 2.2.3.1 discusses the acoustic effects of laryngeal movements.

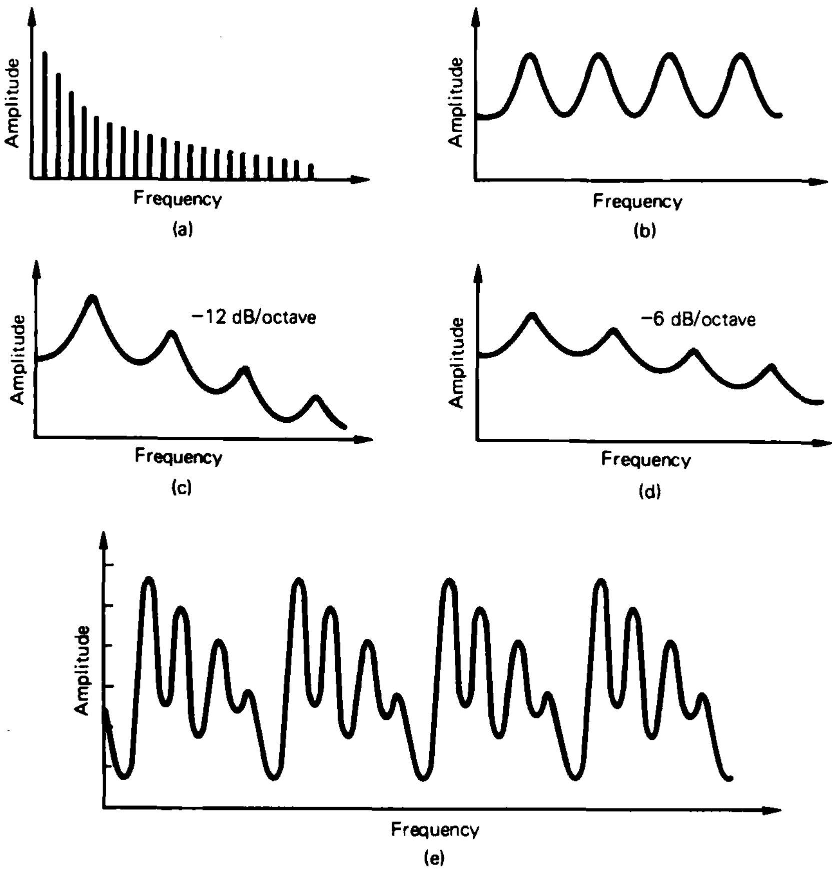

Fig. 1.2 (a) Spectrum of glottal waveform. (b) Resonant response of the vocal tract. (c) Spectral envelope from filtering of glottal source waveform by vocal tract. (d) Boost of 6 dB/octave in spectral envelope of radiated sound wave. (e) Time-domain waveform corresponding to (d). After Clark and Yallop (1990).

changes may occur in the formants. The mode of vocal cord activity also may slightly affect formant centre frequencies and bandwidths. Nevertheless, vowel quality remains constant despite dramatic changes in the source spectrum, because the envelope of the output spectrum undergoes little alteration. The source–filter theory of vowel production therefore suggests that invariants for the solution of the constancy problem reside in the spectral envelope of the lip output. The theory also points towards differences between spectral output envelopes as the principal basis for vowel categorization. These suggestions dovetail with the contemporary treatment of the auditory system as a type of frequency analysis device.

1.3 Physical representations of vowels

1.3.1 Time-domain and frequency-domain representations

This brief review of source–filter theory has already introduced two forms of stimulus representation, time-domain and frequency-domain. Figure 1.3 contains examples obtained with a Kay DSP-5500 spectrograph at the Phonetics Laboratory at the University of Oxford. The upper right panel in Fig. 1.3(A) shows a time-domain waveform for a production of [i] by a male speaker of received pronunciation (RP) English. The corresponding panels in Fig. 1.3(B) and (C) display time-domain waveforms for [ɑ] and [u], respectively. The upper left panels in Fig. 1.3(A)–(C) display frequency-domain representations for each time-domain waveform. The frequency-domain representations are 1024-point fast Fourier transforms (FFTs) taken over the part of each time-domain waveform that is lightly printed in the figure. Both types of representation display changes across the different vowels. In particular, overall peaks appear at different places in the FFTs, representing the effects of altered vocal-tract resonances. The individual lines in the FFTs represent the harmonics of the voicing source. A harmonic does not necessarily occur at the centre frequency of a vocal-tract formant. Therefore, attempts to estimate formant centre frequencies from the harmonics with the largest amplitudes in an FFT are prone to significant errors. Such estimates may be off by $F0/2$, where $F0$ is the momentary fundamental frequency of the voice. To solve this problem, FFT spectra can be smoothed. Cepstral smoothing is often used for this purpose (see, for example, Clark and Yallop 1990).

1.3.2 Speech spectrograms

The single FFTs in Fig. 1.3. naturally cannot represent two vowels of the same quality but of different durations. The durations of vowels do vary with the rate of speech. Even more importantly, durational differences are phonemically distinctive in languages such as Hungarian and Japanese. In these languages, two vowels that endow words with different meanings may have virtually identical spectral properties but different durations. Since speech production always keeps articulators on the move, vocal-tract resonances must constantly change. This fact makes it desirable to represent spectral information as a function of time. Such a depiction would automatically give information on duration.

The spectrogram provides one solution to this problem. Spectrograms display frequency-domain information as a function of time. This representation automatically gives information about vowel duration. The lower panels in Fig. 1.3(A)–(C) show broadband spectrograms for [i], [ɑ], and [u]. The

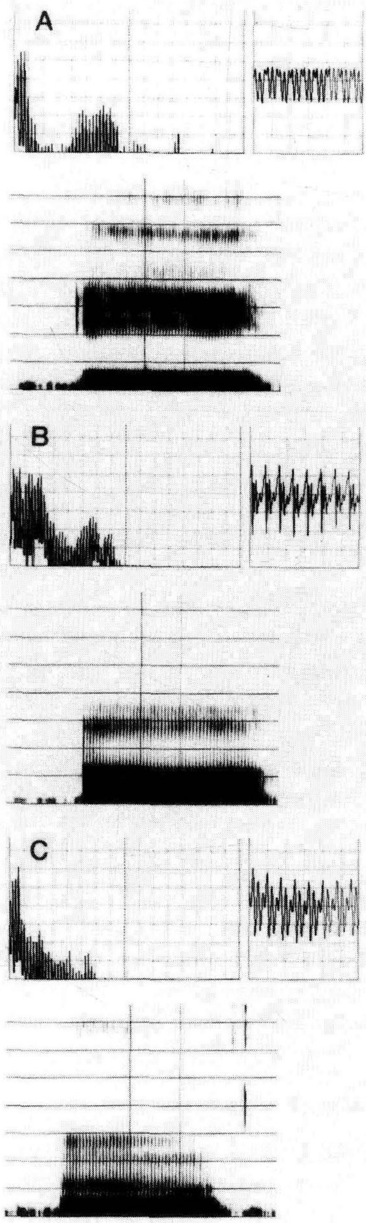

Fig. 1.3 (A) Vocalized vowel [i] produced by a male speaker of RP English. Upper right panel shows time-domain waveform; time calibration marks, 10 ms. Upper left panel shows 1024-point FFT, taken over part of time-domain waveform that is lightly printed; frequency calibration marks, 500 Hz. Lower panel shows broadband spectrogram (300 Hz bandwidths); frequency calibration marks, 1 kHz. (B) Same as (A) for vocalized vowel [ɑ]. (C) Same as (A) for vocalized vowel [u].

time-domain waveforms in each part of Fig. 1.3 were begun at the heavy dotted vertical line laid over each spectrogram. In the spectrograms, changes are visible over time in the heavily blackened frequency regions that represent formants. (In keeping with current practice, we also will use the term 'formant' to refer to such a peak in a frequency-domain representation.) The vertical striations in the spectrograms mark cycles of vocal-fold activity. Narrowband spectrograms, which are not illustrated, would contain narrow quasihorizontal lines. Each line would correspond to an individual harmonic of the glottal source. Concurrent variations in the vertical positions of these lines over time would reflect changes in vocal-fold vibration frequency.

Figure 1.4 portrays whispered versions of the three RP voiced vowels of Fig. 1.3. Figure 1.4 is organized exactly like Fig. 1.3. The FFTs have overall peaks in the same frequency regions as those in Fig. 1.3. Those peaks are better defined in the FFTs for the whispered vowels. In the two figures, energy bands also appear in corresponding places in the broadband spectrograms. The formants are harder to define in the broadband spectrograms for the whispered vowels.

One limitation of the spectrogram is its relatively small dynamic range. This property often makes it hard to obtain accurate readings of formant centre frequencies, much less bandwidths, from spectrograms. A running frequency-domain transform in a three-dimensional space overcomes this limited dynamic range. In such a waterfall display, the *x*- and *y*-axes show frequency and intensity, while the *z*-axis represents time. Successive spectral sections, which are usually smoothed, march along the *z*-axis. Formants therefore appear as ridges whose positions change slowly with time. Examples appear in the papers by Tufts *et al.* (1976) and by Searle *et al.* (1980). One drawback of waterfall displays is that the results are hard to correlate with a speaker's output over long stretches of time. This type of correlation is more easily made with a speech spectrogram. Furthermore, the details in waterfall displays on dynamic changes in output have not proven particularly useful. Spectrograms remain more popular than waterfall displays.

1.3.3 Linear prediction

Linear predictive coding (LPC) is a computational method for estimating formant centre frequencies and bandwidths from digitized samples of the time-domain waveform of a vowel. This form of analysis provides better estimates of formant parameters than does the spectrogram. The method depends on the structure of the time-domain waveform of an epoch of speech and on the assumption of complete, mutual independence of source and filter. The waveform of a voiced vowel has a more or less regular structure over any short portion of the vowel's duration. Therefore, the value of a given sample from the waveform can be predicted, albeit with some error, from the values of *n* of

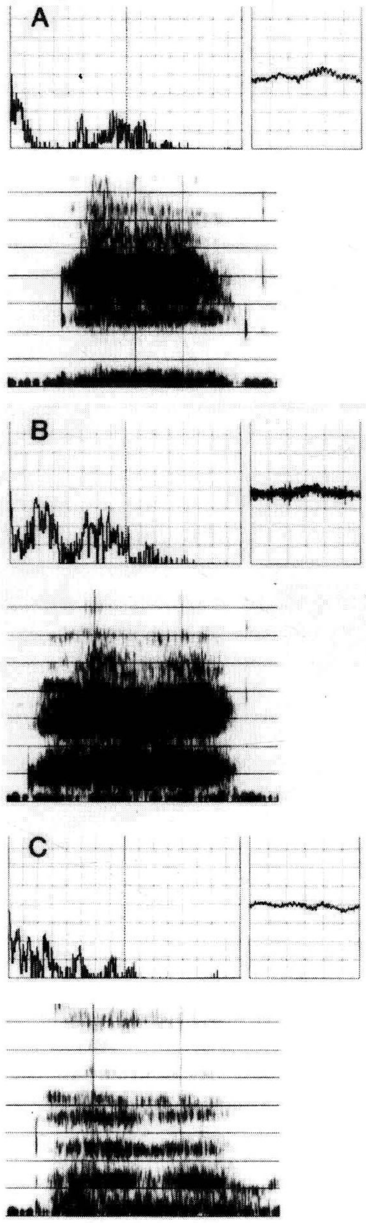

Fig. 1.4 (A) Whispered vowel [i] produced by a male speaker of RP English. Display and calibrations as in Fig. 1.3. (B) Same as (A) for whispered vowel [ɑ]. (C) Same as (A) for whispered vowel [u].