Processing of spoken CVCs in the auditory periphery. I. Psychophysics

Oded Ghitza
AT&T Bell Laboratories, Acoustics Research Department, Murray Hill, New Jersey 07974

(Received 17 February 1993; revised 30 July 1993; accepted 2 August 1993)

This study provides a quantitative measure of the accuracy of the auditory periphery in representing prespecified time-frequency regions of initial and final diphones of spoken CVCs. The database comprised word pairs that span the speech space along Jakobson *et al.*'s binary phonemic features [Tech. Rep. No. 13, Acoustic Laboratory, MIT, Cambridge, MA (1952)]. The time-frequency domain was divided into "tiles" by splitting the frequency range into three bands ([0,1000], [1000,2500], [2500,4000] Hz), and by marking the phonemic time landmarks of the CVC utterance. Fourteen modified versions of this database were generated by introducing well-defined distortions into the time-frequency tiles (or combination of tiles). The performance of eight listeners was measured for each of these versions by using a one-interval two-alternative forced-choice paradigm, to minimize the role of cognition. The results demonstrate that in the first and the second frequency bands, the diphone information is far more important than the consonant information or the vowel information alone. As for the third band, most of the information of the diphone is contained in the consonantal time interval. These observations are common to both the initial and the final consonants of spoken CVCs. The study also provides a direct mapping between Jakobson *et al.*'s features and particular regions in the time-frequency domain. Voicing and nasality are strongly correlated with the diphone information in the first frequency band, graveness and compactness with the diphone information in the second frequency band, and sibilation with the consonantal time interval in the third frequency band. Sustention is equally correlated with the diphone information in the second and the third frequency bands. Since the role of cognition was neutralized to a large extent, the results may also be interpreted as a map of the phonemic distinctive features to some peripheral auditory functions that operate on the corresponding time-frequency regions.

PACS numbers: 43.72.Ar, 43.71.Cq, 43.66.Lj

INTRODUCTION

Stationary intervals rarely occur in fluent speech. Steady-state vowels, for example, usually serve as targets for the articulators. In a wideband spectrogram (like the one shown in Fig. 1) the dynamic nature of speech is reflected in the movement of the formants. Consider, for example, the sequence f–E–l in Fig. 1. The formants of the E sound follow a certain trajectory that is affected by the surrounding sounds f and l. The formant locations change with time and only momentarily reach the prototype values of the vowel E, somewhere during the vowel occurrence.

Numerous studies have been conducted to understand the rules by which (dynamic) speech is either produced (by the articulatory system, e.g., Coker, 1976) or perceived (by the auditory system, e.g., Delattre *et al.*, 1955 or Klatt, 1989). This paper belongs to the second category, where the focus is on how human subjects perceive diphones of spoken CVCs. More precisely, the interest is in measuring how accurate is the auditory representation of diphones of spoken CVCs, in isolation from the cognitive aspects of speech perception. This interest is the first step in a long-term program aimed at understanding how auditory nerve activity is integrated over intervals of 50–150 ms, and over diphones in particular. We assume the existence of peripheral auditory functions that operate on different sections of the auditory nerve. Due to the tonotopic organization of the auditory nerve fibers, the nature of these auditory functions, as well as their accuracy of performance, can be inferred from psychophysical data on the perception of acoustic information contained in different time-frequency regions. Obtaining this psychophysical data is the subject of the present paper.

We used a database that spans the speech subspace associated with initial and with final diphones, and defined along Jakobson *et al.*'s phonemic dimensions (Jakobson *et al.*, 1952). [The database comprises two parts, Voiers' old DRT database for initial diphones (Voiers, 1983), and a new database, also designed by Voiers and termed DALT for Diagnostic ALliteration Test, for final diphones (Voiers, 1991).] In order to relate the phonemic domain to the time-frequency domain, 14 modified versions of the combined database were generated by introducing well-defined distortions into preselected time-frequency regions. The modified versions were tested using Voiers' psychophysical discrimination task. The responses of the subjects were displayed as a distribution along Jakobson *et al.*'s phonemic dimensions, yielding 14 error distributions. Each error distribution provides a quantitative measure of the importance of the particular time-frequency region that was modified.

 0001-4966/93/94(5)/2507/10/$6.00

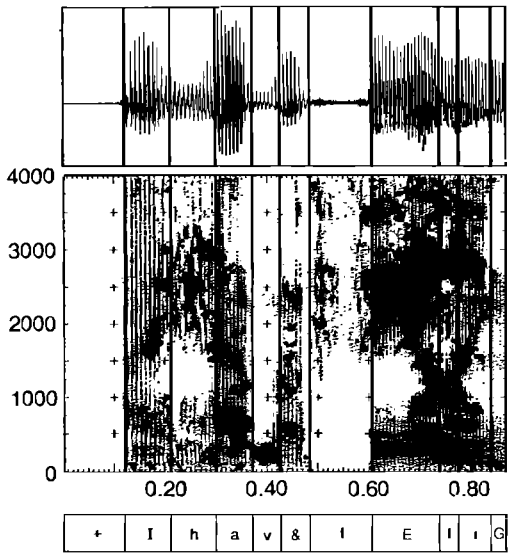

FIG. 1. The waveform and the wideband spectrogram of the sentence "I have a feeling" spoken fluently by a male speaker. Notice, for example, the nonstationarity of the vowel [E].

In Sec. I, the DRT and the DALT are briefly described. In Sec. II, the considerations in defining the 14 modification conditions are discussed. Section II also describes the signal processing procedures that were used to produce the modified DRT and DALT databases. The resulting error patterns are plotted and discussed in Sec. III. Finally, in Sec. IV we draw conclusions from the data.

I. REVIEWING VOIERS' DRT AND DALT

The DRT (Diagnostic Rhyme Test) and the DALT (Diagnostic ALliteration Test) were suggested by Voiers (Voiers, 1983, 1991) as a way of measuring the intelligibility of processed speech. The tests are appropriate for our needs for two reasons. First, the database spans the speech subspace associated with initial and with final diphones. And, second, it allows us to separate the effects of the auditory periphery from those due to cognition.

From acoustic point of view, Voiers' DRT database covers initial diphones of spoken CVCs, and his DALT database covers the final diphones. Tables I and II show the lists of words in the DRT and the DALT, respectively. Each list consists of 96 pairs of confusable words spoken in isolation. Each word is of the CVC type, and words in a pair differ only in their initial (for the DRT) or their final (for the DALT) consonants. The diphones are equally distributed among six phonemic distinctive features (16 word-pairs per feature) and among eight vowels. Both DRT and DALT are defined over the same phonemic features and the same vowels. The feature classification follows the binary system suggested by Jakobson *et al.* (1952).[1] The vowels are [ee] and [it] (high-front), [eh] and [at] (high-back), [oo] and [oh] (low-front), and [aw] and [ah] (low-back).

The psychophysical procedure is also very carefully controlled. The listeners are well trained and are very familiar with the database, including the voice quality of the individual speakers. The experiment is a one-interval two-alternative forced-choice (1I2AFC) experiment. First, the subject is presented visually with a pair of rhymed words. Then, one word of the pair (selected at random) is presented aurally and the subject is required to indicate which of the two words was played. This procedure is repeated until all the words in the database have been presented. The errors can be displayed either in terms of a confusion matrix (between consonants), or as a distribution among the six phonemic distinctive features.

The controlled nature of the database and of the test procedure is the basis for our assumption that all cognitive information needed for the discrimination task is available to the listener prior to the aural presentation (of course, we also assume that the subject is indeed utilizing all this information). If this assumption is correct, an error in identifying the word is due mainly to inaccuracy in the internal auditory representation of the stimulus. Hence, the error list provided by the test reflects errors in the internal human auditory representation during the discrimination task.

TABLE I. Stimulus words used in the DRT (borrowed from Voiers, 1983).

VOICING voiced–unvoiced	NASALITY nasal–oral	SUSTENTION sustained–interrupted
veal–feel	meat–beat	vee–bee
beam–peen	need–deed	sheet–cheat
gin–chin	mitt–bit	vill–bill
dint–tint	nip–dip	thick–tick
zoo–Sue	moot–boot	foo–pooh
dune–tune	news–dues	shoes–choose
voal–foal	moan–bone	those–doze
goat–coat	note–dote	though–dough
zed–said	mend–bend	then–den
dense–tense	neck–deck	fence–pence
vast–fast	mad–bad	than–Dan
gaff–calf	nab–dab	shad–chad
vault–fault	moss–boss	thong–tong
daunt–taunt	gnaw–daw	shaw–chaw
jock–chock	mom–bomb	von–bon
bond–pond	knock–dock	vox–box

SIBILATION sibilated–unsibilated	GRAVENESS grave–acute	COMPACTNESS compact–diffuse
zee–thee	weed–reed	yield–wield
cheep–keep	peak–teak	key–tea
jilt–gilt	bid–did	hit–fit
sing–thing	fin–thin	gill–dill
juice–goose	moon–noon	coop–poop
chew–coo	pool–tool	you–rue
Joe–go	bowl–dole	ghost–boast
sole–thole	fore–thor	show–so
jest–guest	met–net	keg–peg
chair–care	pent–tent	yen–wren
jab–dab	bank–dank	gat–bat
sank–thank	fad–thad	shag–sag
jaws–gauze	fought–thought	yawl–wall
saw–thaw	bond–dong	caught–taught
jot–got	wad–rod	hop–fop
chop–cop	pot–tot	got–dot

TABLE II. Stimulus words used in the DALT (borrowed from Voiers, 1991).

VOICING voiced–unvoiced	NASALITY nasal–oral	SUSTENTION sustained–interrupted
teethe–teeth	screen–screed	seethe–seed
jib–gyp	ring–rig	dish–ditch
mood–moot	noon–nude	goof–goop
brogue–broke	moan–mode	both–boat
liege–leash	gleam–glebe	chief–cheep
ridge–rich	rim–rib	give–gib
prove–proof	tomb–tube	soothe–sued
loathe–loath	gloam–globe	jove–job (e)
led–let	hen–head	rev–reb
have–half	lam–lab	calve–cab
jaws–joss	brawn–broad	froth–fraught
hodge–hotch	nom–nob	slav–slob
peg–peck	gem–jeb	flesh–fletch
lathe–lath	fan–fad	path–pat
flaws–floss	long–log	frothe–fraud
fob–fop	bomb–bob	bosh–botch
SIBILATION sibilated–unsibilated	**GRAVENESS** grave–acute	**COMPACTNESS** compact–diffuse
peach–peak	sheave–sheathe	league–lead
kiss–kith	skim–skin	hick–hit
sues–soothe	rufe–ruth	fugue–feud
poach–poke	oaf–oath	bloke–bloat
breeze–breathe	neap–neat	creek–creep
bridge–brig	miff–myth	sling–slim
truce–truth	rube–rude	luke–loop
clothes–clothe	lobe–load	rogue–robe
bess–beth	deaf–death	mesh–mess
badge–bag	dab–dad	lag–lad
ross–wroth	shawm–shawn	dog–daub
notch–knock	top–tot	cock–cot
ledge–leg	web–wed	egg–ebb
mass–math	raff–rath	knack–nap
maws–mothe	trough–troth	gong–gone
bodge–bog	sauve–swathe	chock–chop

II. MODIFYING DIPHONES IN THE DRT AND DALT DATABASE

In this section we shall define the modifications imposed on the DRT and the DALT word pairs. The definitions hold for both the DRT and the DALT databases. Although the illustrations (examples and figures) relate to modifications of initial diphones in the DRT database, their analogous counterparts are appropriate for illustrating modifications of final diphones in the DALT database.

A. Defining a tile

Figure 2 shows the waveforms and the wideband spectrograms of the DRT word-pair shock/mock spoken by the same male speaker. Also indicated are the boundaries between the phonemes. As shown in the figure, the speech signal is bandlimited to 4000 Hz. Since the words differ only in their initial diphone, the main difference between the spectrograms is in the time-frequency region associated with that diphone, i.e., the region bounded by the bold lines containing the initial consonant (either [S] or [m]) and the left, coarticulated part of the vowel [a].

Figure 3 shows a diagram of the time-frequency region

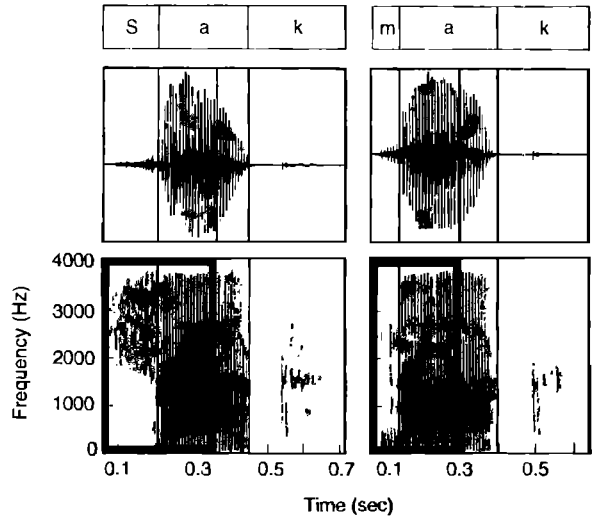

FIG. 2. The waveforms and the wideband spectrograms of the DRT word-pair shock/mock spoken by a male speaker. Also indicated are the boundaries between the phonemes. Since the words differ only in their initial diphone, the main difference between the spectrograms is in the time-frequency region associated with that diphone, i.e., the region bounded by the bold lines containing the initial consonant (either [S] or [m]) and the left, coarticulated part of the vowel [a].

occupied by a spoken CVC word, like the word shock or mock of Fig. 2. The time-frequency region of the initial diphone is subdivided into six "tiles." The frequency boundaries (from the bottom up) are 0, 1000, 2500 Hz and the highest frequency in the band, say 4000 Hz. The time landmarks are (from left to right) the beginning of the word ($t=0$), the transition from the initial consonant to the vowel ($t=t_{tr}$) and the midpoint of the vowel ($t=t_{mid}$). For stop consonants, t_{tr} is the transition from the stop release to the vowel.

Definition 1: Let the tiled region of Fig. 3 be viewed as a 3 by 2 matrix, with three rows (the frequency bands [0,1000], [1000,2500], [2500,4000], in Hz) and two col-

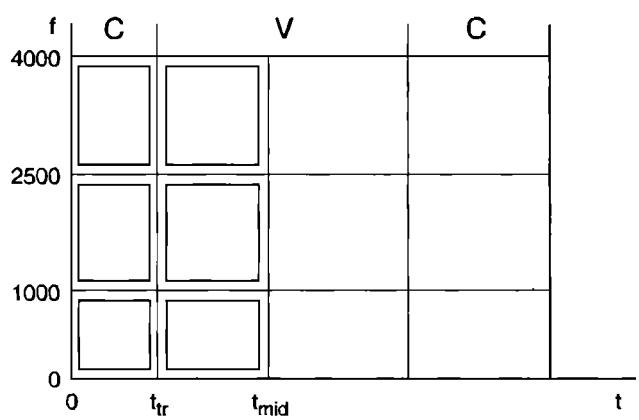

FIG. 3. A diagram of the time-frequency domain occupied by a spoken CVC word. The time-frequency region of the initial diphone is sub divided into six "tiles." The frequency boundaries (form the bottom up) are 0, 1000, and 2500 Hz and the highest frequency in the band, say 4000 Hz. The time landmarks are (from left to right) the beginning of the word ($t=0$), the transition from the initial consonant to the vowel ($t=t_{tr}$) and the midpoint of the vowel ($t=t_{mid}$). For stop consonants, t_{tr} is the transition from the stop release to the vowel.

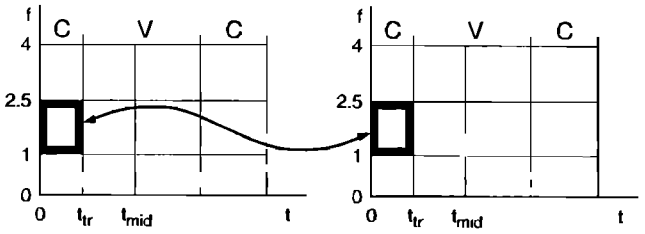

FIG. 4. A diagram of the time-frequency domain occupied by a prototype DRT word pair, where the region corresponding to the initial diphones are divided into six tiles each. Illustrated is the interchange operation $I_{2.C}$.

umns ($[0,t_{tr}]$, $[t_{tr},t_{mid}]$). Then, tile $T_{i.j}$ is the tile containing the ith frequency band of the jth time band, with $i \in \{1,2,3\}$ for the first, second, and third frequency band, respectively, and $j \in \{C,V\}$.

For example, $T_{2.C}$ is the tile containing the second frequency band ([1000,2500], in Hz) of the initial consonant ($[0,t_{tr}]$). To allow simultaneous selection of more than one tile, we will extend the notation in an obvious way. $T_{i.CV}$ will represent the ith frequency band of the whole diphone, and $T_{12.j}$ (or $T_{123.j}$) will represent the first and second frequency bands combined (or the full frequency band) of the jth time band.

The choice of the frequency boundaries was motivated by two observations. First, by analogy between measured auditory nerve responses of the cat (e.g., Delgutte and Kiang, 1984; Sachs and Young, 1979) and possible auditory nerve responses of a human, we expect a significant difference between the properties of the firing patterns of low CF (say, below 1000 Hz) and high CF fibers (CF stands for characteristic frequency. It indicates the place of origin of a nerve fiber along the basilar membrane). At low CFs harmonics are resolved, and neural discharges of auditory nerve fibers are phase locked to the underlying driving component (i.e., synchrony is maintained). At high CFs, frequency resolution is poor and synchrony of neural discharges is greatly reduced. Obviously, there is no distinct boundary between these auditory nerve regions. Rather, the change in properties is gradual. However, our working hypothesis is that the region of transition is around 1000 Hz. The choice of the next boundary (at 2500 Hz) was motivated by the importance of the movements of the second formant, in speech perception. The second frequency band (from 1000 to 2500 Hz) is, therefore, taken to be the range of the second formant (Peterson and Barney, 1952).

B. Defining an interchange operation

Figure 4 shows a diagram of the time-frequency domain occupied by a prototype DRT word pair, where the regions corresponding to the initial diphones are divided into six tiles each.

Definition 2: Let W^1 and W^2 be a DRT word pair, and let $T^k_{i.j}$ be the tile $T_{i.j}$ of word W^k, $k=1,2$. An interchange $I_{i.j}$ between words W^1 and W^2 is the interchange of tiles $T^1_{i.j}$ and $T^2_{i.j}$.

The example in Fig. 4 illustrates the interchange $I_{2.C}$ between the two words. Of course, one can interchange

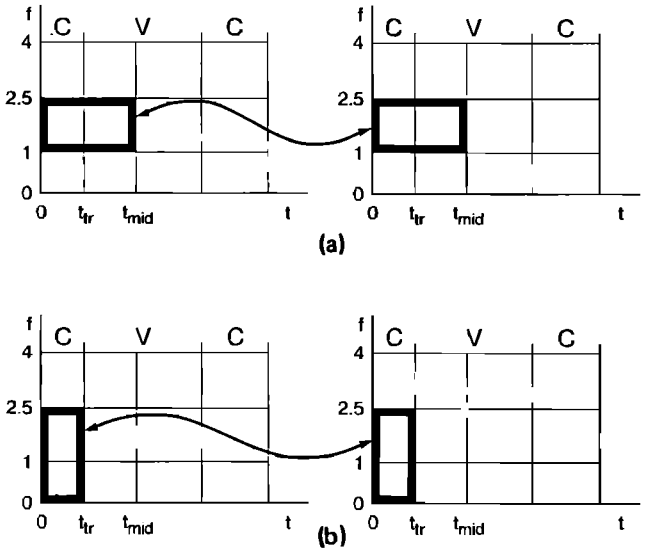

FIG. 5. An illustration of interchange operations (a) $I_{2.CV}$ and (b) $I_{12.C}$.

more than one tile at a time. Again, we will use the notation $I_{2.CV}$ to indicate the interchange of the second frequency band of the whole diphone [Fig. 5(a)], $I_{12.C}$ to indicate the interchange of the first and second bands of the consonant alone [Fig. 5(b)], etc. Listed in Table III are the 14 different interchange operations that were used in the experiment.

C. Motivation for interchange operation

Rather than using the interchange operation, other ways of modifying the DRT words could be considered. We tested, for example, the possibility of removing a tile (or tiles) from the two words in the pair, leaving the corresponding time-frequency region empty. Thus subjects would have to respond to stimuli that do not contain the information carried by the time-frequency region of interest. This kind of operation, however, resulted in speech that sounded unnatural. Rather, it sounded like a superposition of frequency bands that are out of time alignment. The outcome of the psychophysical experiment was, therefore, biased since subjects are trained to listen to natural, although degraded, speech.

TABLE III. Interchange operations used in this study.

	Interchange	Freq. band (Hz)	Time interval
1	$I_{1.C}$	[0,1000]	$[0,t_{tr}]$
2	$I_{1.V}$	[0,1000]	$[t_{tr},t_{mid}]$
3	$I_{1.CV}$	[0,1000]	$[0,t_{mid}]$
4	$I_{2.C}$	[1000,2500]	$[0,t_{tr}]$
5	$I_{2.V}$	[1000,2500]	$[t_{tr},t_{mid}]$
6	$I_{2.CV}$	[1000,2500]	$[0,t_{mid}]$
7	$I_{3.C}$	[2500,4000]	$[0,t_{tr}]$
8	$I_{3.V}$	[2500,4000]	$[t_{tr},t_{mid}]$
9	$I_{3.CV}$	[2500,4000]	$[0,t_{mid}]$
10	$I_{12.C}$	[0,2500]	$[0,t_{tr}]$
11	$I_{12.V}$	[0,2500]	$[t_{tr},t_{mid}]$
12	$I_{12.CV}$	[0,2500]	$[0,t_{mid}]$
13	$I_{123.C}$	[0,4000]	$[0,t_{tr}]$
14	$I_{123.V}$	[0,4000]	$[t_{tr},t_{mid}]$

We also tested the possibility of removing a tile (or tiles) from the two words in the pair and filling the corresponding time-frequency region with white Gaussian noise (keeping the energy of the noise and the energy of the signal over this region equal). Indeed, this kind of operation resulted in speech that was perceived natural (although degraded). However, for some of the words this operation resulted in speech that was perceived as a word outside the word pair. For example, in words that contain voiced stop consonants, the resulting stimulus after modification could be a word with an unvoiced stop consonant (e.g., the word *bone* in the DRT pair bone/moan could be perceived as *pone*). This, again, violates the conditions of the DRT psychophysical procedure since it cannot be considered any longer as a 1I2AFC experiment.

These trials led us to choose the interchange operation of definition 2. Here, a modification results in a word that belongs to the word-pair under modification, although degraded. As we shall describe next, the signal processing procedure was designed to keep unnatural sounding artifacts to a minimum.

D. Signal processing procedure

The database comprises the DRT and the DALT word lists spoken by three male speakers (speakers RH, CH, and JE from Voiers' recordings). An experienced phonetician segmented by hand all the words in the database, using a multidimensional display that included the waveform, its wideband spectrogram and an aural input (via headphones). All time landmarks (including the midpoints of the phonemes) were marked. For signal processing, the database was low-pass filtered to 3600 Hz and sampled at an 8-kHz rate. Each word was filtered to provide three bands ([0,1000], [1000,2500], and [2500,3600] Hz). Each filter was an FIR with 161 taps. Fifteen modified versions of the database were created, a baseline version and 14 "tiled" versions. The baseline version is simply the summation of the three bands. There is some difference between an original stimulus and its corresponding baseline stimulus. The baseline version served as our reference (rather than the original itself) since the tiled versions are created by using the same bandlimited signals.

The tiled versions were synthesized by cutting and pasting the appropriate bandlimited signals, as defined by the interchange operations listed in Table III. For the cut and paste, the hand-segmented time landmarks were used. To demonstrate the interchange operation, let us denote the waveform by $s^k(t)$, and the length of the word W^k belonging to a word pair by T^k with $k=1,2$. Let $s_i^k(t)$ be the bandlimited signal corresponding to the ith frequency band of $s^k(t)$. Let $\hat{x}(t)$ be the tiled version of $x(t)$, resulting from an interchange operation. Let $\Delta t_V^k = t_{mid}^k - t_{tr}^k$, where t_{tr}^k is the time of transition from the consonant to the vowel, t_{mid}^k is the midpoint of the vowel and, therefore, Δt_V^k is the duration of the left, coarticulated part of the vowel of W^k. Finally, let $\Delta t_V = (\Delta t_V^1 + \Delta t_V^2)/2$ be the average duration of the left, coarticulated part of the vowel, averaged over W^1 and W^2.

1. Interchanging a tile in a diphone

If, for example, the first frequency band is to be interchanged over the entire diphone, the following steps are performed. First, the interval $[0,t_{tr}^1+\Delta t_V]$ is removed from $s_1^1(t)$. Although the hand segmented midpoint t_{mid}^1 was replaced by $t_{tr}^1+\Delta t_V$, we still consider the time interval $[0,t_{tr}^1+\Delta t_V]$ to be the duration of the diphone, assuming that the difference between the vowel durations Δt_V^1 and Δt_V^2 is small. Next, the interval $[0,t_{tr}^2+\Delta t_V]$ is cut from $s_1^2(t)$. This portion is then added to the time interval $[t_{tr}^1+\Delta t_V, T^1]$ of $s_1^1(t)$, with the "stitching" point at $t_{tr}^1+\Delta t_V$, resulting in the tiled signal $\hat{s}_1^1(t)$. Two points are noteworthy. First, the overall length of $\hat{s}_1^1(t)$ is different from T^1 because usually $t_{tr}^1 \neq t_{tr}^2$, and, second, in order to smooth the discontinuity at the stitching point, the cutting at $t_{tr}^1+\Delta t_V$ and at $t_{tr}^2+\Delta t_V$ [which are the cutting points for $s_1^1(t)$ and for $s_1^2(t)$, respectively] is accompanied by a taper over a 20-ms-long window centered at these points. In the final step of the interchange procedure the desired tiled version of $s^1(t)$ is created by simply summing the corresponding bandlimited signals: i.e., $\hat{s}^1(t)=\hat{s}_1^1(t)+s_2^1(t)+s_3^1(t)$. The tiled version of $s^2(t)$ is created in an analogous way.

2. Interchanging a tile in a consonant

The procedure is similar to the above except that the intervals $[0,t_{tr}^1]$ and $[0,t_{tr}^2]$ are interchanged, and the smoothing window is 10 ms long.

3. Interchanging a tile in a vowel

Again, the procedure is the same except that the intervals $[t_{tr}^1,t_{tr}^1+\Delta t_V]$ and $[t_{tr}^2,t_{tr}^2+\Delta t_V]$ are interchanged. Also, the smoothing window is 10 ms long at the consonant cutting points and 20 ms long at the vowel cutting points.

To mask remaining stitching artifacts, low level noise was added to all 15 versions. We used Gaussian noise with

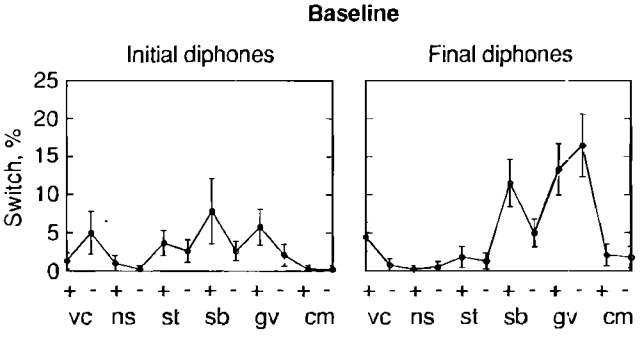

FIG. 6. The average and the 95% confidence interval for the baseline versions of the DRT (left) and DALT (right). The abscissa of every plot indicates the 12 phonemic categories: "vc" is for voicing, "ns" for nasality, "st" for sustention, "sb" for sibilation, "gv" for graveness, and "cm" for compactness. The "+" sign stands for attribute present and the "—" sign for attribute absent. The ordinate is termed "switch," and it represents the number of words in the category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener "switched" to the opposite category). The switch is represented in percents, relative to 16 (the total number of words per phonemic category).

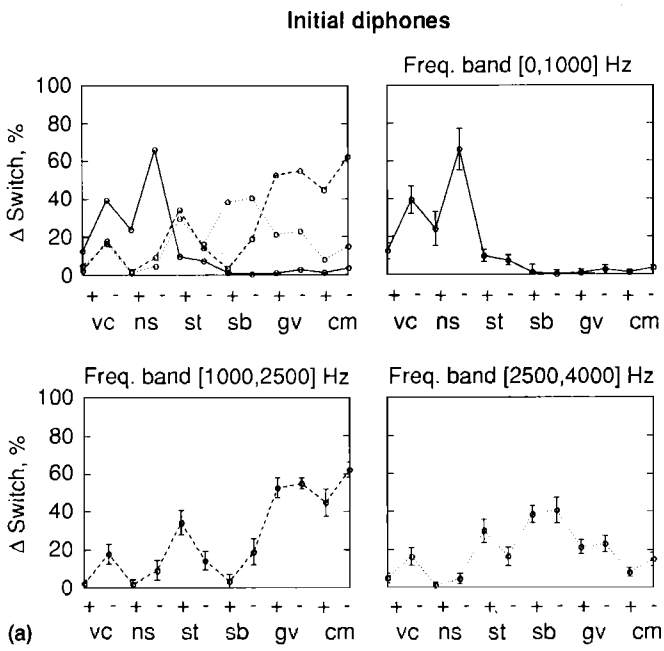

FIG. 7(a). Human performance under interchange of each frequency band over the entire diphone, on the DRT database. The upper left plot is a summary of the other three plots, with the confidence-interval bars omitted. The abscissa is as in Fig. 6. The ordinate is termed *Δswitch*, since it represents the *additional* number of switches, relative to the baseline version, that occurred due to the particular interchange operation. Note that the line connecting the measurements is only for display purposes, to enable the reader to distinguish between patterns that belong to a particular interchange condition. The upper right plot shows the amount of *Δswitch*, in percent, under interchange operation $I_{1,CV}$. The lower left plot is for $I_{2,CV}$, and the lower right plot is for $I_{3,CV}$. Notice that voicing and nasality are strongly correlated with the first frequency band of the diphone, graveness and compactness with the second frequency band of the diphone, and sibilation with the third frequency band of the diphone.

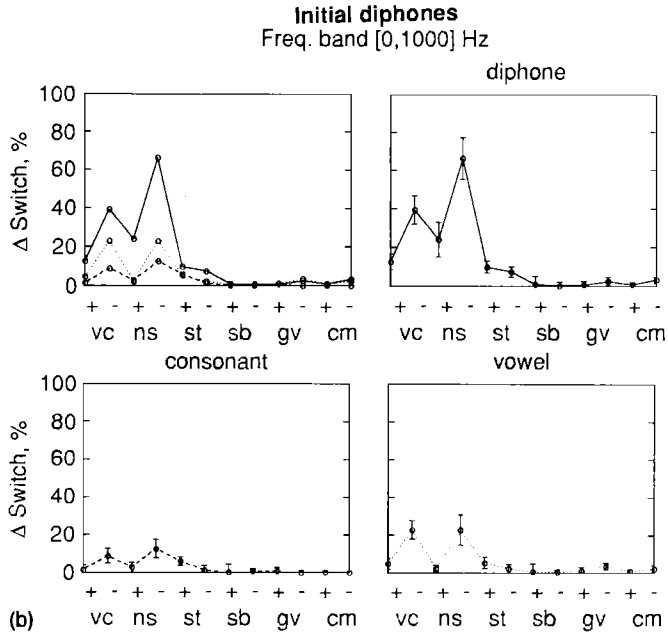

FIG. 7(b). A focus on the first frequency band. The relative importance of an interchange over the entire diphone is compared to an interchange over the consonant or the vowel alone (in our notation, we compare interchange operations $I_{1,CV}$, $I_{1,C}$, and $I_{1,V}$). The upper right plot is the same as the upper right plot of (a). Figure legend is as in (a). Notice that the diphone information is far more important than the consonant information or the vowel information alone.

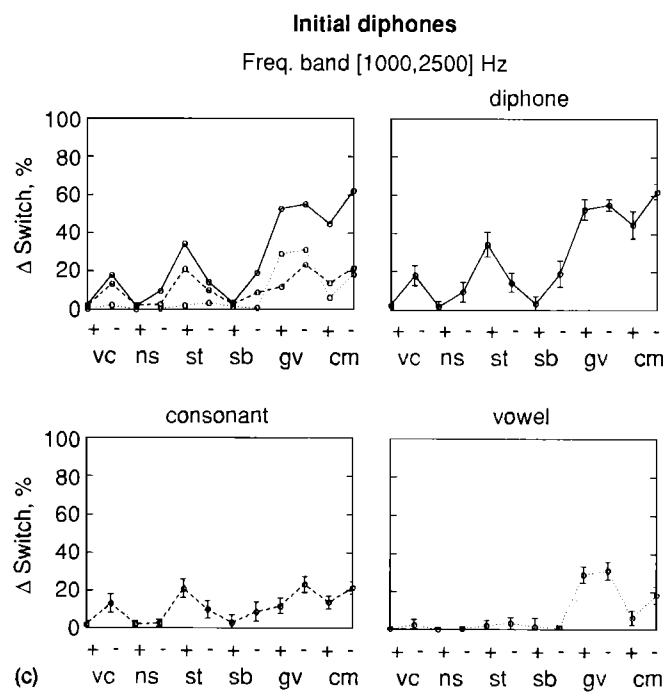

FIG. 7(c). A focus on the second frequency band. The relative importance of an interchange over the entire diphone is compared to an interchange over the consonant or the vowel alone (in our notation, we compare interchange operations $I_{2,CV}$, $I_{2,C}$, and $I_{2,V}$). The upper right plot is the same as the lower left plot of (a). Figure legend is as in (a). Notice that the diphone information is far more important than the consonant information or the vowel information alone.

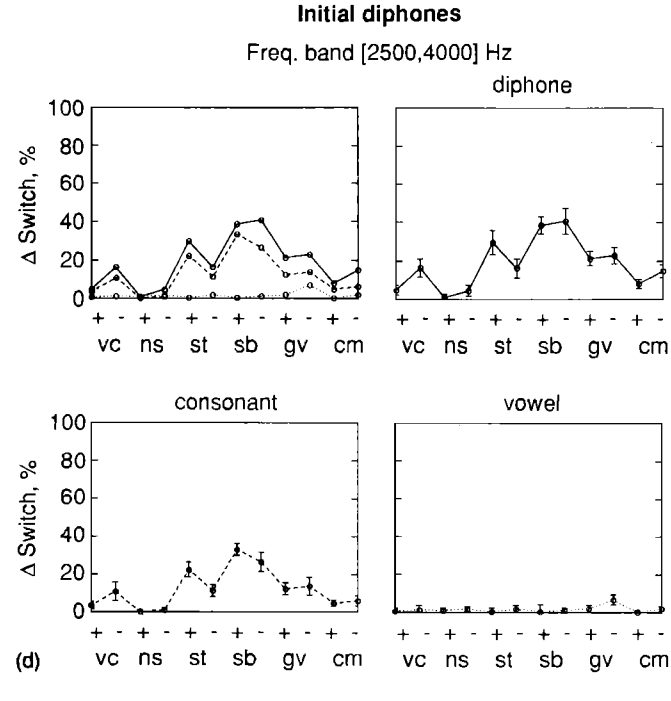

FIG. 7(d). A focus on the third frequency band. The relative importance of an interchange over the entire diphone is compared to an interchange over the consonant or the vowel alone (in our notation, we compare interchange operations $I_{3,CV}$, $I_{3,C}$, and $I_{3,V}$). The upper right plot is the same as the lower right plot of (a). Figure legend is as in (a). Notice that most of the information of the diphone is contained in the consonantal time interval.

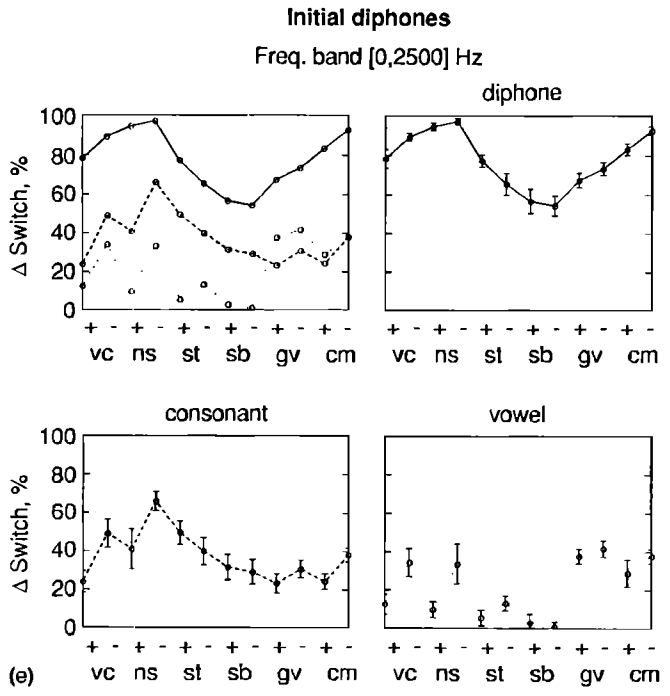

FIG. 7(e). The perceptual importance of the diphone in the first and second frequency bands combined (i.e., [0,2500] Hz), compared to the perceptual importance of the consonantal part or the vowel part alone (i.e., $I_{12.CV}$, $I_{12.C}$, and $I_{12.V}$). Figure legend is as in (a).

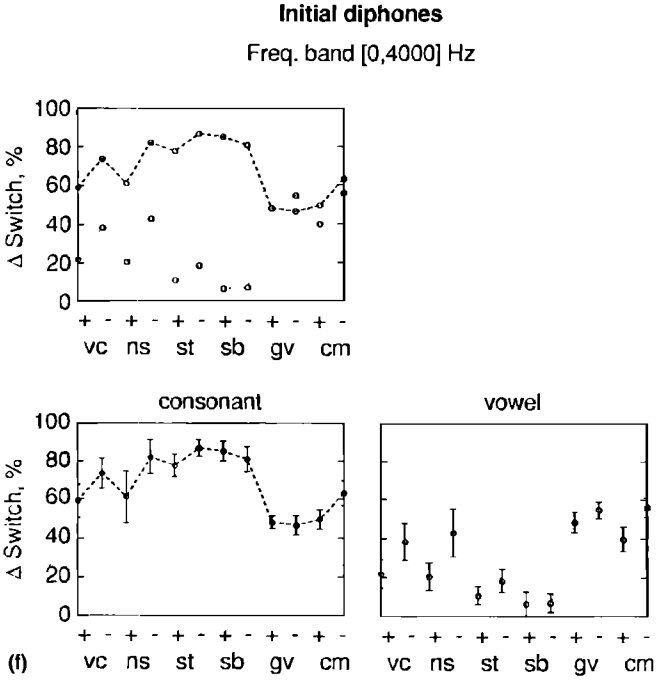

FIG. 7(f). The perceptual importance of the consonantal and the vowel parts in the full frequency band [0,4000) Hz (i.e., $I_{123.C}$ and $I_{123.V}$). We did not test the performance under a complete diphone interchange (i.e., $I_{123.CV}$) assuming that it will cause a complete categorical switch. Figure legend is as in (a).

a power spectral density that follows the long-time power density spectrum of continuous speech. The signal-to-noise ratio (SNR) was 30 dB.

All versions were sent to Dynastat Inc. (a company established by Voiers) for the psychophysical evaluation. To comply with Dynastat's procedure, the processed words were recorded at a rate of one word every 1.3 s. For the recordings to sound continuous over time, we first set the energy of the noise generator to a level that remained unchanged until all the words in the DRT word list had been recorded in sequence. To record a particular word, the signal was amplified (or attenuated) by a gain factor that was calculated in advance to maintain the desired SNR.

III. EXPERIMENTAL RESULTS

We tested all 15 versions (baseline, and the interchange operations listed in Table III). We define a "DRT run" as an experimental run of a particular version of the DRT word list, spoken by the three speakers. "DALT run" is defined in an analogous way. Overall, therefore, we performed 30 experimental runs. For each run, the raw data were organized in the form of a matrix with 12 columns and 24 rows. Each column represents a phonemic category (attribute present or attribute absent for each of the six phonemic features—see Tables I and II) and each row represents a listener–speaker pair [eight listeners, each performed the experiment three times (once for each of three speakers) giving 24 rows]. A matrix element indicates the number of errors for the ith listener–speaker pair, over all 16 words that represent the jth phonemic category (i.e., we averaged across all vowels). Therefore, a matrix element can be any integer between 0 and 16.

Calculating statistics across the rows, we computed the average and the 95% confidence interval for every column (i.e., for every phonemic category we averaged across all vowels, all listeners, and all speakers).

A. RESULTS

The results are plotted in Figs. 6–8. Figure 6 shows the resulting average and the associated confidence interval for the baseline versions of the DRT (left) and DALT (right). The abscissa of every plot indicates the 12 phonemic categories: "vc" is for voicing, "ns" for nasality, "st" for sustention, "sb" for sibilation, "gv" for graveness, and "cm" for compactness. The "+" sign stands for attribute present and the "−" sign for attribute absent. The ordinate is termed "switch," and it represents the percentage of words in the category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener "switched" to the opposite category).

Figure 7 summarizes the results of the other 14 DRT runs. Panels (a)–(f) show comparisons of various groups selected from the 14 conditions. Each panel shows four plots, where the upper left plot is a summary of the other three plots, with the confidence-interval bars omitted. For each plot, the abscissa is as in Fig. 6. As for the ordinate, from the measured switches for the current interchange condition, we subtracted the number of switches for the baseline version (Fig. 6, left). We termed the ordinate "Δ switch," since it represents the *additional* number of switches, relative to the baseline version, that occurred due

to the particular interchange operation. Note that the line connecting the measurements is only for display purposes, to enable the reader to distinguish between error patterns that belong to a particular interchange condition.

Figure 7(a) summarizes the performance of the listeners under interchange of each frequency band over the entire diphone. The upper right plot shows the amount of $\Delta switch$, in percent, under interchange operation $I_{1.CV}$, (i.e., when the frequency band [0,1000] Hz of the initial diphone is interchanged). The lower left plot is for $I_{2.CV}$ (i.e., when the frequency band [1000,2500] Hz of the initial diphone is interchanged), and the lower right plot is for $I_{3.CV}$ (i.e., the frequency band [2500,4000] Hz of the initial diphone is interchanged). Several observations are noteworthy. First, voicing and nasality are strongly correlated with the first frequency band of the diphone, graveness and compactness with the second frequency band of the diphone, and sibilation with the third frequency band of the diphone. Sustention is equally correlated with the second and the third frequency bands. Second, under $I_{1.CV}$, words that belong to voicing and nasality demonstrate an asymmetrical $\Delta switch$ across the "+" and "−" attributes. For example, a word from the "−" category of nasality will switch to the "+" category with probability of 0.66. However, a word from the "+" category of nasality will switch to the "−" category with probability of only 0.24. In con-

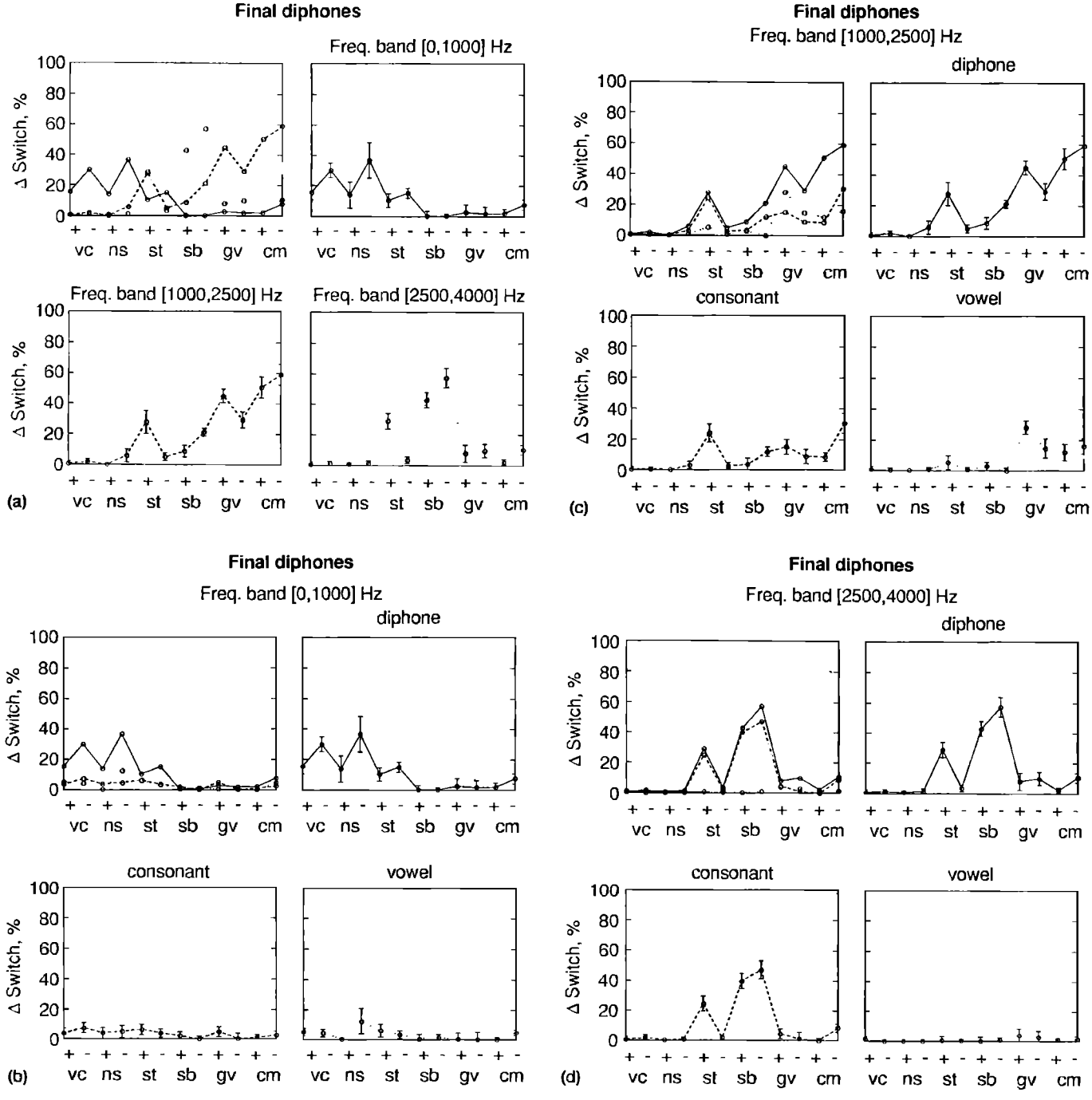

FIG. 8. (a) Same as Fig. 7(a), for DALT. (b) Same as Fig. 7(b), for DALT. (c) Same as Fig. 7(c), for DALT. (d) Same as Fig. 7(d), for DALT. (e) Same as Fig. 7(e), for DALT. (f) Same as Fig. 7(f), for DALT.

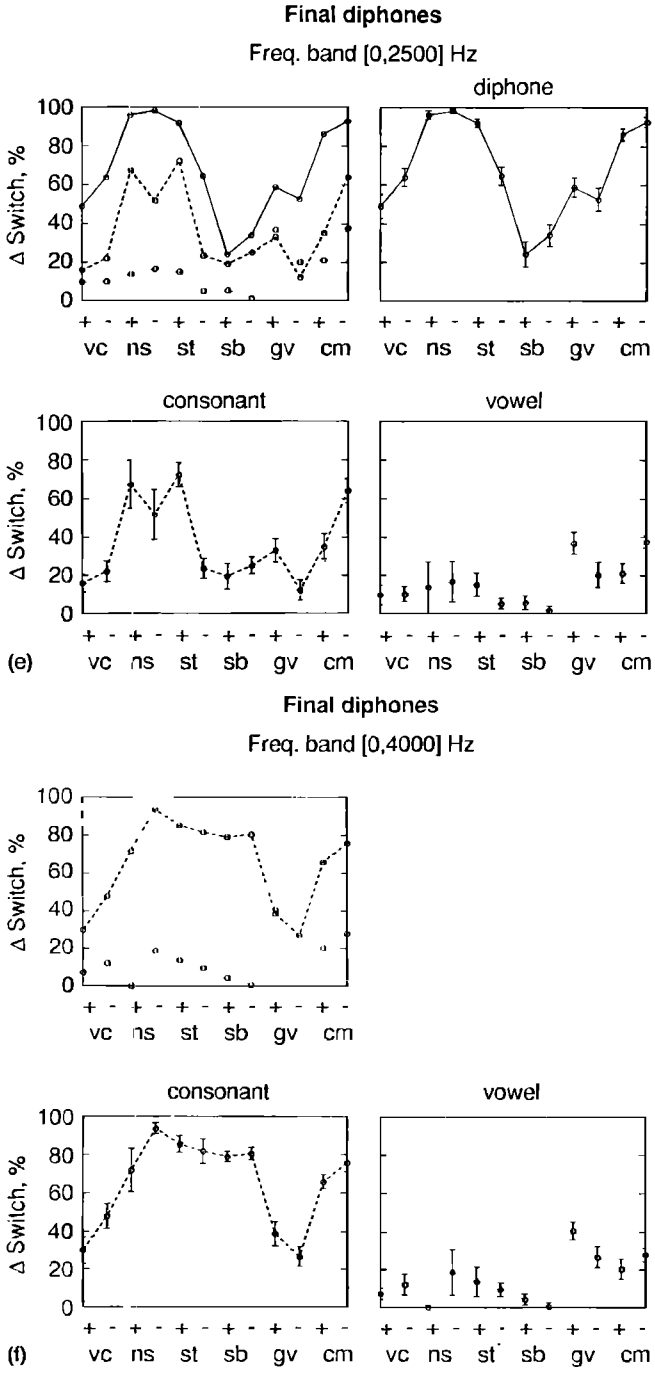

FIG. 8. (*Continued.*)

trast, a symmetrical Δ*switch* occurs under $I_{2.CV}$, for words that belong to graveness and compactness. Δ*switch* is also symmetrical under $I_{3.CV}$, for words that belong to sibilation.

In Fig. 7(b) we focus on the first frequency band and examine the relative importance of an interchange over the entire diphone compared to an interchange over the consonant or the vowel alone (in our notation, we compare interchange operations $I_{1.CV}$, $I_{1.C}$, and $I_{1.V}$). Note that the upper right plot is the same as the upper right plot of Fig. 7(a). The Δ*switch* under $I_{1.CV}$ is much greater than the Δ*switch* under $I_{1.C}$ or $I_{1.V}$ alone. This demonstrates that in the first frequency band, the diphone information is far more important than the consonant information or the vowel information alone.

In Fig. 7(c) we focus on the second frequency band and again examine the relative importance of an interchange over the entire diphone compared to an interchange over the consonant or the vowel alone (in our notation, we compare interchange operations $I_{2.CV}$, $I_{2.C}$, and $I_{2.V}$). The upper right plot is the same as the lower left plot of Fig. 7(a). Again, the Δ*switch* under $I_{2.CV}$ is much greater than the Δ*switch* under $I_{2.C}$ or $I_{2.V}$ alone. This demonstrates that in the second frequency band too, the diphone information is far more important than the consonant information or the vowel information alone.

In Fig. 7(d) we focus on the third frequency band and examine the relative importance of an interchange over the entire diphone compared to an interchange over the consonant or the vowel alone (in our notation, we compare interchange operations $I_{3.CV}$, $I_{3.C}$ and $I_{3.V}$). The upper right plot is the same as the lower right plot of Fig. 7(a). The Δ*switch* under $I_{3.V}$ is negligible. Rather, the Δ*switch* under $I_{3.C}$ alone is very close to the Δ*switch* under $I_{3.CV}$. This demonstrates that in the third band, most of the information of the diphone is contained in the consonant time interval.

Figure 7(e) shows the perceptual importance of the diphone in the first and second frequency bands combined i.e., [0,2500] Hz), compared to the perceptual importance of the consonantal part or the vowel part alone (i.e., $I_{12.CV}$, $I_{12.C}$, and $I_{12.V}$).

Figure 7(f) shows the perceptual importance of the consonantal and the vowel parts in the full frequency band [0,4000] Hz (i.e., $I_{123.C}$ and $I_{123.V}$). We did not test the performance under a complete diphone interchange (i.e., $I_{123.CV}$) assuming that it will cause a complete categorical switch.

Figure 8 summarizes the results of the 14 DALT runs. The same comparisons are made as in Fig. 7 and analogous notations are used. Somewhat surprisingly, the qualitative behavior is similar to that of Fig. 7. However, there are quantitative differences, i.e., in the numerical values of the Δ*switch* in various conditions.

IV. DISCUSSION

In this study we provide a relationship between articulatory/phonemic dimensions and particular regions in the time-frequency domain. A series of psychophysical experiments was conducted for this purpose. A few of these experimental conditions have been studied earlier, mainly in the context of the perception of synthetic speech sounds (e.g., Cooper *et al.*, 1952; or Delattre *et al.*, 1955). Most of the experimental conditions, however, have not been addressed before and are focused on measuring the importance of pre-specified time-frequency regions for the perception of spoken CVCs.

From the outset, we assumed the existence of distinct, albeit unknown, auditory functions which operate on different time-frequency regions. From this perspective, the most general result of this study is the demonstration that time-frequency regions that are occupied by a diphone play

a central role in the perception of CVCs. In the first and the second frequency bands, a diphone interchange is far more dominant than a consonant interchange or a vowel interchange alone. In the third band, a diphone interchange is as dominant as the interchange of the consonant alone.

The assumption of the existence of distinct auditory functions is reinforced by the second important finding of this study: There is a direct relationship between Jakobson *et al.*'s phonemic features and specific time-frequency regions of a diphone. Diphone information in the first frequency band is strongly correlated with voicing and nasality, diphone information in the second frequency band with graveness and compactness, and the consonantal time interval in the third frequency band with sibilation. Sustention is equally correlated with the diphone information in the second and the third frequency bands [such a behavior is expected, since the acoustic manifestation of "interrupt" (which is one of the binary attributes of sustention) is an abrupt temporal change, spread over a wide frequency range]. Evidence for a correlation of this nature was reported previously. Applying multidimensional scaling techniques on Miller and Nicely's confusion matrices (1955)[2], Shepard (1972) and Wish and Carroll (1974) demonstrated that place and manner phonemic features (which are strongly coupled to Jakobson *et al.*'s features) do relate to some, unspecified, perceptual dimensions. Applying similar techniques on the same database, Soli and Arabie (1979) inferred a relationship between the place and manner features and certain acoustic properties of speech. Those studies could provide only indirect evidence because of the limitations associated with the multidimensional scaling technique. In contrast, our study provides direct evidence of the relationship between the articulatory domain (expressed in Jakobson *et al.*'s terms) and the time-frequency domain.

Another difference between those studies and ours is noteworthy. Shepard, Wish and Carroll, and Soli and Arabie based their analyses on Miller and Nicely's experimental data which, as a consequence of the experimental procedure, reflected the performance of the overall auditory system—peripheral and cognitive parts combined. The perceptual dimensions that emerge from their analyses are, therefore, associated with central parts of the auditory pathway as well. In contrast, the experimental procedure used in the DRT and DALT is a 1I2AFC, in which the role of cognition is reduced to the largest extent possible. Therefore, the observed relationship between the phonemic features and the time-frequency domain may be interpreted as a map of particular phonemic features to *peripheral* auditory functions that operate on the corresponding time-frequency regions of the diphone. This relationship may be used to define the task each auditory function must perform, as well as the required degree of accuracy.

[1]According to Jakobson *et al.* (1952), the voicing feature characterizes the nature of the source, being periodic or nonperiodic. The nasality indicates the existence of a supplementary resonator. The terms sustention and sibilation are due to Voiers. They correspond respectively to the continuant-interrupted and strident-mellow contrasts of Jakobson *et al.* Finally, graveness and compactness represent broad resonance features of the speech sound, related to place of articulation.

[2]Miller and Nicely measured perceptual confusions among 16 English consonants. They used 16-CV stimuli (with the consonants followed by the vowel [a] (as in father) spoken by five female speakers over different noise, low-pass and high-pass conditions. As for the psychophysical paradigm, a subject was first presented aurally with the stimulus and then was forced to indicate which of the 16 consonants was played (i.e., an identification procedure).

Coker, C. H. (**1976**). "A model of articulatory dynamics and control," Proc. IEEE **64**, 452–460.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 579–606.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Delgutte, B., and Kiang, N. Y. -S. (**1984**). "Speech coding in the auditory nerve: I. Vowel-like sounds," J. Acoust. Soc. Am. **75**, 866–878.

Jakobson, R., Fant, C. G. M., and Halle, M. (**1952**). "Preliminaries to speech analysis: the distinctive features and their correlates," Tech. Rep. No. 13, Acoustic Laboratory, MIT, Cambridge, MA.

Klatt, D. H. (**1989**). "Review of selected models of speech perception," in *Lexical Representation and Process,* edited by W. Marslen-Wilson (MIT, Cambridge, MA), pp. 169–226.

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Sachs, M. B., and Young, E. D. (**1979**). "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," J. Acoust. Soc. Am. **66**, 470–479.

Shepard, R. N. (**1972**). "Psychological representation of speech sounds," in *Human Communication: A Unified View,* edited by E. E. David and P. B. Denes (McGraw-Hill, New York), pp. 67–113.

Soli, S. D., and Arabie, P. (**1979**). "Auditory versus phonemic accounts of observed confusions between consonant phonemes." J. Acoust. Soc. Am. **66**, 46–59.

Voiers, W. D. (**1983**). "Evaluating processed speech using the Diagnostic Rhyme Test," Speech Technol. **1**(4), 30–39.

Voiers, W. D. (**1991**). "Effects of noise on the discriminability of distinctive features in normal and whispered speech," J. Acoust. Soc. Am. **90**, 2327 (A).

Wish, M., and Carroll, J. D. (**1974**). "Applications of individual differences scaling to studies of human perception and judgment," in *Handbook of Perception, Vol. II,* edited by E. C. Carterette and M. P. Friedman (Academic, New York), pp. 449–491.