On the rise and fall of Spanish diphthongs

Elisabeth Mauder and Vincent J. van Heuven

0. Introduction

When, in Spanish or in any other language, two full vowels assume abutting positions across a syllable boundary, the natural tendency towards an alternating CV-CV structure may be restored by reducing one of the two full vowels to a semi-vowel, thereby creating a single diphthong instead of a sequence of two full vowels. The resulting diphthong depends on the relative sonority of the two full vowels involved and the position of the more sonorous vowel within the pair.

Open vowels such as /a/ have greater inherent intensity (all else being equal) than closed vowels such as /i/. Peterson & Lehiste (1959) report a difference of 5.5 dB greater intensity for American English /a/ than for /i/. If the two vowels in a VV-sequence such as /a-i/ were pronounced with equal effort, the intensity envelope of the sequence would be falling from the first to the second vowel; in the VV-sequence /i-a/, on the other hand, the intensity contour would be rising. In the present paper we will use the terms "rising" and "falling" vowel sequences in the above sense.

A diphthong is a sequence of a vowel V and a semi-vowel (or glide) G within a single syllable. Naturally, the inherent intensity (or: sonority) of semi-vowels is weaker than that of full vowels. As before, then, falling diphthongs comprise a sequence of a full vowel followed by a glide VG (e.g. /aj/ and oj/ in English *fine* and *boy*, respectively); by the same token, rising diphthongs consist of a leading glide element followed by a full vowel (e.g. French /jɛ̃/ and /wa/ in *rien* and *roi*, respectively)[1]. The reader should bear in mind that the falling/rising distinction refers to the development of intensity over the course of the vowel sequence; it does not refer to the closing (rising) versus opening (falling) movement of the tongue during the diphthong (for a fuller introduction to the terminology cf. Jones 1918: §§ 219-224).

If reduction to diphthong takes place, there are two possibilities, corresponding to the two possible sequences of an open and an non-open vowel:

---

[1] The second element of a falling diphthong is not necessarily a semi-vowel. In so-called breaking diphthongs the low-intensity element is a schwa, often but not inherently, a surface reflection of an underlying /r/ (cf. Br. Eng. /fɪə/ *fear* /kɛə/ *care*, etc.). Regular alternation of falling and rising diphthongs is found in Frisian (Cohen, Ebeling & van Holk 1978) where singular nouns have falling (breaking) diphthongs (/dɔəs/ 'box', /bɪən/ 'leg') but their inflected forms have the corresponding rising versions (/dwaskə/ 'small box', /bjɛnə/ 'legs').

Linguistics in the Netherlands 1996, 171–182. DOI 10.1075/avt.13.17mau
ISSN 0929–7332 / E-ISSN 1569–9919

/aV/ and /Va/. In the former case, the second V reduces to G, creating a falling diphthong, in the former case, the first V-element reduces, yielding a rising diphthong. Apparently, there is a general constraint in Spanish that excludes the open vowel /a/ from reduction to glide. Any non-open vowel, however, can be reduced to G in a VV-sequence (Navarro Tomas, 1932; Gili Gaya, 1966)[2].

The possibility for reduction of VV-sequences to diphthong in Modern Standard Spanish is crucially constrained by two conditions which involve word stress:

(1) *Enable diphthongization*: When word stress falls on a vowel outside the VV-sequence, or on the more sonorous (i.e. more open) of the two adjacent vowels, the VV-sequences is almost invariably reduced to a diphthong in careless speech (such that the less open vowel reduces to its corresponding glide).

(2) *Disable diphthongization*: When the stress is on the less open (i.e. less sonorant) of the two adjacent vowels, reduction to diphthong is blocked.

There are, however, sources of exception to the second constraint: Some Latin American dialects allow diphthongization of sequences where the less open vowel is stressed in the Standard Spanish version of the word. This is reported for Chile (Rabanales 1960) and Bolivia (van Wijk 1961) and is evident in almost any text of Argentinean 'gaucho-literature'. This leads, for example, to the pronunciation of the word *país* /pa'is/ 'country' as /'pajs/ or *maestro* /ma'estro/ 'master' as /'majstro/. The reduction of such sequences, however, involves not only the reduction of a non-open vowel to a glide but also a change in the word's stress pattern: the reduction depends on a shift of the word stress from the less sonorous to the more sonorous vowel; the conditioning stress-shift must precede the reduction to diphthong, since only unstressed vowels can be reduced to a glide.

In the gaucho dialect, where the frequency of the reduction process can be investigated, one finds, however, a striking asymmetry between rising and falling VV-sequences: in those cases where the less sonorous vowel is stressed, the reduction of falling VV-sequences (such as /aí > áj/) occurs on a much larger scale than reduction of the rising variant (such as /ía > já/), i.e., falling diphthongs are created much more frequently than rising ones.

This outcome is in contrast with the general frequency of rising and falling diphthongs in Spanish, where rising diphthongs are notably more frequent than falling ones (Gili Gaya, 1966). On the other hand, it seems that Spanish, in this respect, is exceptional among the languages of the world: there is general

[2] Note: the VV to diphthong process is simplified for expository reasons here. In fact, in any sequence of two vowels with unequal vowel height specifications, it is invariably the less open vowel that is reduced to glide.

consensus that the falling diphthong is the more frequent type cross-linguistically, and that the rising type is infrequent.

In order to explain the asymmetry in the reduction of VV-sequences in the gaucho dialect, the production and perception of stress in VV-sequences might provide important clues: Considering that the stress-shift is a necessary condition for the reduction of VV-sequences with stress on the less sonorous vowel, differences in the accuracy of stress perception between rising and falling sequences might be a plausible explanation for the more frequent reduction of falling sequences; i.e. if the probability of stress shift were asymmetrical for the two sequences, for mere psycho-acoustic reasons, this might explain the asymmetry in the reduction process, since the reduction occurs almost 'automatically' once the more sonorous vowel is (perceived as) stressed. Should this be the case, two conditions would have to be met:

- the perception of stress position should be less accurate in falling VV-sequences than in rising VV-sequences, i.e., stress position should be more difficult to perceive in /a-i/ than in /i-a/.
- errors in stress perception should not occur randomly; whenever stress is not perceived clearly, the more sonorous vowel should more often be perceived as stressed than the less sonorous one.

If these two conditions were met, there would be a clear phonetic (perceptual) basis for the observed asymmetry.

The present study was set up to find out if the position of word-stress in Spanish is perceived as accurately in rising and falling VV-sequences (in which one V may be realized as a glide G). The experiment comprised an acoustic analysis of tokens of such sequences as well as a perceptual determination of stress position by a group of native listeners.

1. Experiment I: Acoustic analysis

1.1 Materials. We chose stimuli containing sequences of the maximally different vowels [i] and [a] in both rising and falling VV sequences and with stress orthogonally varied over the two positions, in lexical items (*maníaco/maniaco* 'maniac' and *su maíz* 'his corn' / *sumáis* 'you *pl.* add')[3] and a set of nonsense words where all four vowel/stress combinations are embedded in otherwise identical syllables (*coníato / coniáto / conaíto / conáito*).

---

[3] Although the final sibilants in *su maíz* and *sumáis* would be pronounced different (/θ/ vs. /s/) in peninsular Spanish, the contrast does not arize in Latin American Spanish, since /θ/ does not occur there and orthographic *z* is always pronounced as /s/.

The target words were presented in context in order to reduce initial stress bias and final lengthening effects (van Heuven, 1987a; van Heuven & Menert, 1996).

In languages such as English and Dutch the position of the stressed syllable is marked more elaborately (by a conspicuous pitch change, as well as by greater intensity and duration and by spectral expansion) when the target word is in focus (i.e. presented by the speaker as imparting important information, cf. van Heuven, 1994 and references given there); the pitch change is lost (Nooteboom, 1972; van Heuven, 1987; Sluijter, 1995) and some temporal (Sluijter & van Heuven, 1995) and spectral (van Bergem, 1993) reduction is found when the target is out of focus. Since the aim of the experiment is to study errors in the perception of stress position, non-focussed targets were included in the materials to see whether indeed more perceptual errors would be found there. Consequently, two contextual versions were created for each of the eight target words, one with narrow focus (i.e. contrastive accent) on the target word, and a complementary one with contrastive accent somewhere in the word group following the target word:

Q1	*¿Qué*	*le*	*hizo*	*decir*	*otra*	*vez?*	
	'What	him	he-made	say	once	more?	
A1	*Le*	*hizo*	*decir*	*'su*	*maíz'*	*otra vez*	
	'Him	he-made	say	"his corn"		once more'	

Q2	*¿A quién le*	*hizo*	*decir*	*'su maíz'*	?	
	'Who him	he-made	say	"his corn"?'		
A2	*Le hizo*	*decir*	*'su maíz'*	*a*	*Miguel*	
	'Him he-made	say	"his corn"	by	Michael	

The set of 16 stimulus expressions was read three times by a male native speaker of Chilean Spanish, a professional performer, and recorded in a sound-proofed cabin at the Phonetics Laboratory of Leiden University[4].

1.2 Acoustic analysis. Of all tokens only the answer sentences were digitally stored (10 KHz, 12 bits, 4,5 KHz LP) and subjected to a Robust LPC formant analysis (formants F1 through F5, and associated bandwidths B1 through B5, 256 point window with 100 point time shift, Willems, 1987), and to pitch (F0) extraction by the method of subharmonic summation (Hermes, 1988). Properties within three acoustic domains were chosen for the analysis (cf. van Heuven, 1996):

---

[4] In fact, the complete stimulus material was recorded once more after the speaker had been instructed to speak as fast as he could comfortably manage (assuming that errors in the perception of stress position would arise in fast speech sooner than in normal speech). For reasons of space, the results of this part of the material will not be presented here.

within three acoustic domains were chosen for the analysis (cf. van Heuven, 1996):

- *Temporal domain*: duration of the individual vowels. Beginning and end of the VV-sequence were defined by eye with auditory feedback; the boundary between the two vowels was defined as the temporal mid-point of the F1 movement.
- *Pitch domain*: the excursion size (measured as the F0 interval between the lowest and the highest pitch within the target domain, expressed in semitones, ST) and relative temporal position of the F0 peak within the VV-sequence (expressed as a percentage of the vowel duration, with a negative value for peaks occurring in the first vowel).
- *Spectral domain*: For the determination of vowel quality, the Hertz values of -F1 (as a measure of vowel height) and -F3 (as a measure of vowel backness)[5] were converted to Barks (a scale that reflects the frequency resolution of the human hearing mechanism, cf. van Heuven, 1988 and references given there); a 'theoretical schwa' was defined axiomatically as a point in the -F1 by -F3 plane by taking the mean F3 across all [a]-tokens and the mean F1 across all [a] *and* [i] tokens (figure 1).

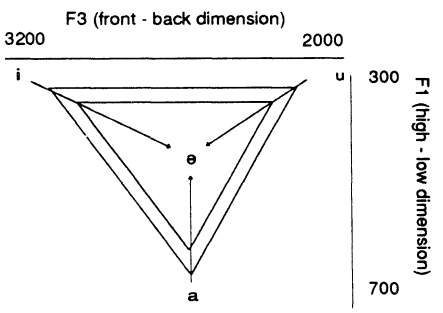

Figure 1. Hypothetical position of reduced (unstressed) and expanded (stressed, accented) vowel tokens in acoustic vowel diagram

1.3 Results. The following figures, which summarize the results, give the mean values for the rising and falling realizations of each pair (accumulated over lexical and nonsense items), for + and - Focus targets separately.
For *vowel duration*, the expectation was that stressed vowels should generally be longer than unstressed ones and that the +focus condition should even increase

[5] Normally, -F2 would be a better acoustic correlate of vowel backness. However, the second formant in our recordings was very weak, possibly due to the presence of nasal antiresonances in the immediate vicinity of the target vowels. Since F3 runs more or less parallel to F2 in the case of non-back vowels, the former measure was adopted instead.

the duration of the stressed vowel, thereby increasing the differences between the stressed and unstressed versions. The results for the duration values are presented in figure 2. It is evident that in general the expectation is borne out: the first vowel, whether /i/ or /a/, is clearly longer in the stressed version compared to the unstressed one. Differences for the second vowel are smaller, and run even counter to the prediction in the case of -focus /i-a/. Prosodic accent on the target word increases the relative duration of the stressed vowel only slightly. As for differences between the rising and falling sequences, it is evident that vowel duration varies more in the falling /a-i/ sequences than in the rising /i-a/ ones. In the latter ones the differences are affected more strongly by the absence of prosodic accent. It is thus the falling sequences where vowel length contains relatively more information as to the position of the stressed vowel.

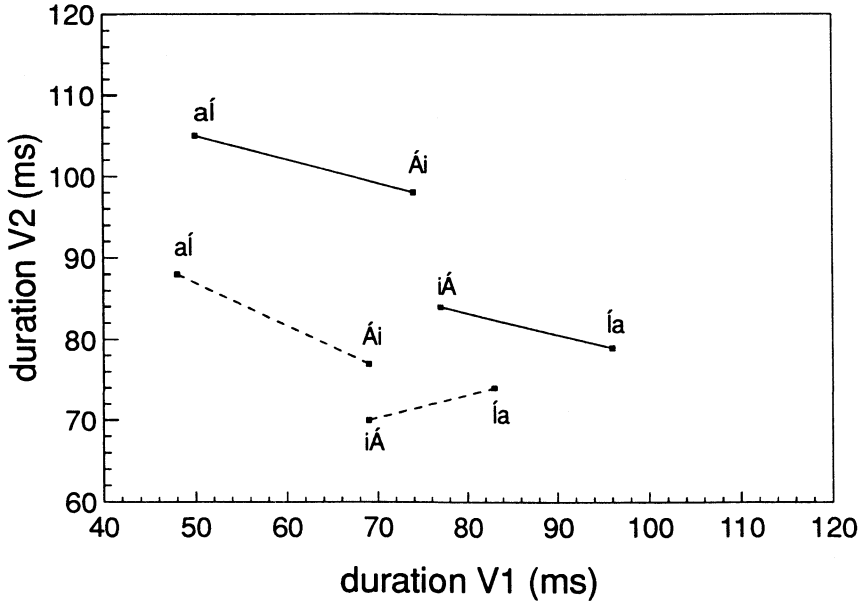

Figure 2. Effects of stress position on duration of first (V1) and second (V2) vowel in rising /i-a/ vs. falling /a-i/ sequences for targets in +focus (sloid lines) and -focus (dotted lines).

For the *F0-movement*, the expectation was that the temporal position of the pitch peak should vary depending on whether the first or second vowel is stressed, but that the excursion should be perceptually insignificant in the -focus condition. Figure (3) shows the effects of internal composition of the vowel sequence and focus condition on the excursion size and temporal location of the pitch peak. When the target sequence is out of focus, the excursion size is small, and quite

probably perceptually negligible (≤ 4 semitones on average). Moreover, the location of the F0-peak within the VV-sequence does not depend on the stress pattern; what we see instead is a well-known effect of so-called intrinsic vowel pitch to the effect that /i/-vowels have a somewhat higher pitch (roughly 20 Hz) than /a/-vowels (all else being equal, cf. Lehiste & Peterson, 1961): the F0-peak invariably lies in the /i/-vowel, whether stressed or not. When the VV-sequence is in focus, however, there is a large accent-lending pitch movement of some 10 semitones. When the first vowel is stressed, the F0-peak is reached at or slightly before the end of the first vowel; when the second vowel is stressed the F0-peak is shifted well into the second vowel. Crucially, the time-shift is about twice as large for falling VV sequences /a-i/ as for rising sequences /i-a/.

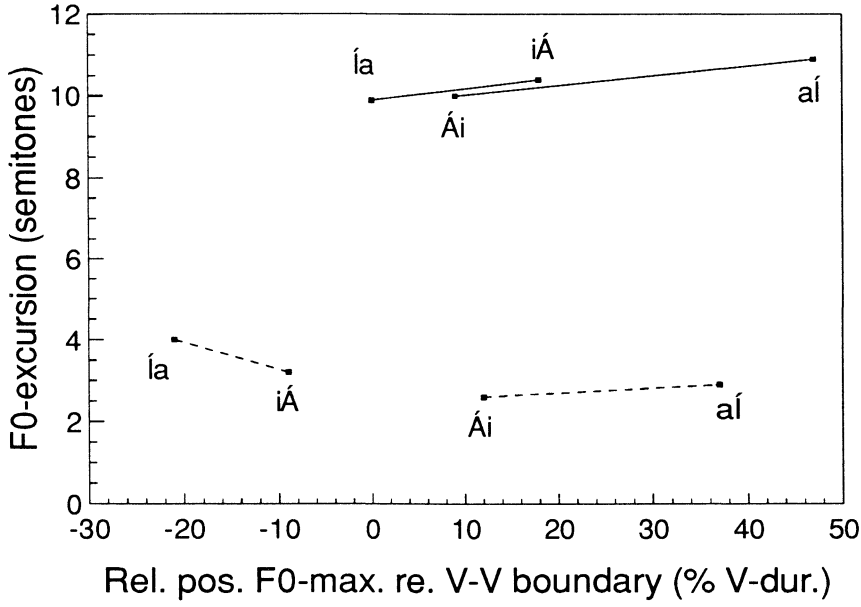

Figure 3. Effects of stress position on location of peak-FO (re. VV-boundary, in percent of vowel duration) and FO excursion size (in semitones) in rising /i-a/ vs. falling /a-i/ sequences for targets in +focus (sloid lines) and –focus (dotted lines).

For *vowel reduction*, there are claims that Spanish vowels hardly change in quality (if at all) between stressed and unstressed conditions (e.g. Navarro Tomas, 1932; Dalbor, 1969; Dauer, 1983; Toledo 1988; Gili Gaya, 1966). On the other hand, in directly adjacent vowels there must be some degree of reduction, and it can be expected that unstressed vowel tokens should be more strongly affected than stressed ones. The results are given in figure 4 (panel A

for falling VV-sequences /a-i/; panel B for rising sequences /i-a/). The figure shows the formant trajectories in the –F1 by –F3 plane (i.e. our acoustical representation of the traditional articulatory height by backness vowel diagram (cf. figure 1); only the left-hand side of the diagram is being shown, since this is the part of the vowel diagram where /a-i, i-a/ trajectories are found.

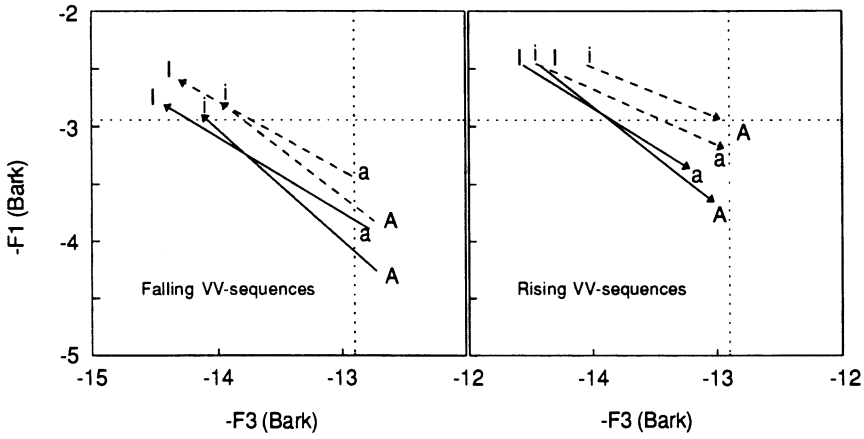

Figure 4. Effect of stress position on vowel quality expansion/reduction of start and end points of formant trajectories for targets in +focus (solid lines) and -focus (dotted lines) in rising sequences /i-a/ (right-hand panel) and falling sequences /a-i/ (left-hand panel). Horizontal axis represents vowel backness; vertical axis represents vowel height. The hypothetical center of the vowel space (schwa) is located at the crossing of the hair lines.

The results for the falling sequences (panel A) are perfectly straightforward. First, the entire trajectory, i.e. both starting points and end points, for sequences pronounced in -focus conditions runs closer to the center of the vowel diagram (represented by the intersection of the hair lines in the figure) than the +focus counterparts. This indicates that there is considerable reduction in Spanish unaccented vowels, in contradistinction to traditional claims in the literature (see above). Second, regardless of focus condition, when the trajectory is from stressed /a/ to unstressed /i/ the starting point is more open for /a/ whilst the /i/ is centralized (backed); when the trajectory runs from unstressed /a/ to stressed /i/ the opposite effect obtains: the /a/ is centralized (raised) whilst the /i/ is expanded (fronted). This is tantamount to saying that the stressed portion of a closing VV-sequences is spectrally expanded whereas the unstressed portion is spectrally reduced.

before. The effect of stress is minimal, but in the predicted direction, for +focus sequences. However, the effect of stress in the –focus trajectories is counter-intuitive in two respects: the degree of reduction for unstressed /i/ is as expected but much larger in –focus than in +focus trajectories; the end points of the /i-a/ trajectories, whether stressed or unstressed, are very close, if not coincident with /ə/. In –focus sequences, then, the effects of stress (shift) in terms of spectral expansion/reduction are less clearly marked than in the +focus counterparts.

1.4 Discussion of the results of the acoustic analysis. In +focus VV-sequences the effects of stress position are considerably better marked for falling /a-i/ sequences than for rising /i-a/ sequences, in terms of peak-F0 position, vowel duration, and vowel reduction/expansion. In the –focus condition, there is no stress-relevant F0-information in either sequence; duration differences remain relatively stable (re. +focus) in rising sequences but decrease in falling sequences; vowel quality differences, finally, are substantial and straightforward for –focus falling diphthongs but unsystematically distributed over the first and second part of the sequences in –focus rising sequences. Our general conclusion so far is that falling sequences are better marked for stress position than rising sequences. For the listener, stress perception should thus be easier in falling sequences, certainly under the +focus condition; in the –focus condition, stress perception should generally be more difficult, as there is no longer any F0-information available for either of the two sequences. The perceptual advantage of falling sequences should therefore diminish in –focus sequences.

2. Experiment II: stress perception

2.1 Method. A perception experiment was carried out with 9 native speakers of Spanish (4 peninsular and 5 Latin American) who took part in the experiment voluntarily. The same material as for the acoustic analysis was used. Stimuli were played twice, in different orders, in a quiet room. The listeners indicated on which vowel (/i/ or /a/) they perceived word stress.

2.2 Results and discussion. Since there were no significant order effects and informants were fairly consistent in their judgment (in 85% of the cases), results will be presented for all listeners collapsed over both orders. Figure 5 shows the results of the perception test in terms of percent error in stress assignment for rising versus falling VV-sequences, separately for errors concerning stressed /a/ and /i/ in +focus (left) and –focus targets (right)[6].

---

[6] Due to the fact that in all stimuli only one of two adjacent vowels is stressed, the perception of any stressed vowel as 'unstressed' implies erroneous perception of 'stress' on the other.

rising versus falling VV-sequences, separately for errors concerning stressed /a/ and /i/ in +focus (left) and –focus targets (right)[6].

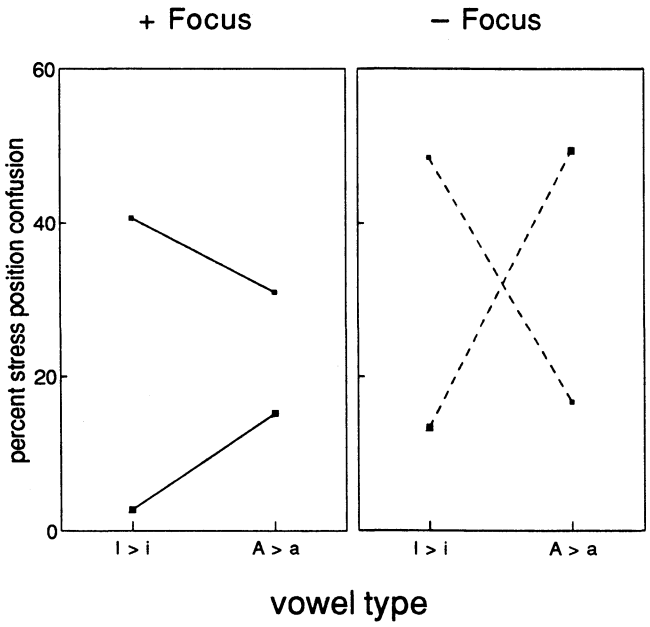

Figure 5. Percent confusions of stressed vowels as unstressed in rising /i-a/ and falling /a-i/ VV-sequences and identity of stressed vowel ('I' vs. 'A'), in targets +focus (left-hand panel) and –focus (right-hand panel).

Overall error rate in stress assignment was on the order of 25%. For sequences in +focus performance is better than for –focus sequences (20 versus 30 % error). As predictable from the acoustical measurements, stress position is, indeed, perceived more often correct in falling than in rising sequences, but only when the target word is accented, i.e. is in focus. Differences between the sequences vanish, however, in the –focus condition. Finally, and crucially, perception of stress is more difficult on /i/ in rising sequences and more difficult on /a/ in falling sequences. This interaction of vowel type and sequence can be interpreted as (either stimulus or response) bias favoring the perception of stress on the second vowel of any VV-sequence. Observe that there is no indication in

[6] Due to the fact that in all stimuli only one of two adjacent vowels is stressed, the perception of any stressed vowel as 'unstressed' implies erroneous perception of 'stress' on the other.

However, there is one interaction in the perception test that does not follow from the acoustic data: the superiority (i.e. perceptual stability) of falling /a-i/ sequences is found only in +focus targets in the perception test, whereas we expected that some measure of superiority would remain in the –focus falling sequences. The latter expectation is not borne out by the perceptual data: stress position is perceived equally poorly in –focus rising and falling sequences. We must assume, therefore, that the perceptually relevant information underlying the superiority of falling sequences lies in the position of the F0-peak within the VV-sequence.

3. Conclusion

The basic prediction for this experiment was that, on the basis of the observed pattern of diphthongization of VV-sequences in some Latin American dialects, word stress should be perceived more accurately in rising vowel sequences than in falling ones and errors in stress perception should systematically favor stress perception on the more sonorous vowel, i.e. /a/ over /i/.

The results of this study do not support either hypothesis: the falling sequences /a-i/ clearly contain more acoustic information about the position of word stress than rising sequences /i-a/. In the absence of prosodic accent on the target word, information as to the position of word stress is severely reduced in both sequences, and errors become equally frequent accordingly.

The reasons for the stress shift in VV-sequences with original stress on the less sonorous vowel and for the higher frequency of reduction of falling sequences in the Argentinean gaucho dialect must therefore be sought elsewhere. We will refrain from making any concrete suggestions at this time.

The results of our experiments, however, are in line with the observation (cf. §0) that falling diphthongs are the more popular type in the languages in the world. If, for whatever reason, a monophthongal vowel destabilizes into a diphthong, our data correctly predict that the falling type will be the preferred choice on account of the greater perceptual stability of this type over the rising alternative.

4. References

Bergem, D. van (1993) Acoustic vowel reduction as a function of sentence accent, word stress, and word class on the quality of vowels, *Speech Communication*, 12, 1-23.

Cohen, A. C.L. Ebeling, P. Eringa, K. Fokkema and A.G.F. van Holk (1978) *Fonologie van het Nederlands en het Fries*, Martinus Nijhoff, Den Haag.

Dalbor, J.B. (1969) *Spanish Pronunciation: Theory and Practice*, Holt, Rinehart & Winston, New York.

Dauer, R.M. (1983) Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics*, 11, 51-62.

Gili Gaya, S. (1966) *Elementos de Fonética General*, Editorial Gredos, Madrid.

Hermes, D.J. (1988) Measurement of pitch by subharmonic summation, *Journal of the Acoustical Society of America*, 83, 257-264.

Heuven, V.J. van (1987a) An unusual effect on the perception of stress, *Proceedings of the 11th International Congress of Phonetic Sciences*, Estonian Academy of Sciences, S.S.R., Tallinn, Vol.V. 306-308.

Heuven, V.J. van (1987b) Stress patterns in Dutch (compound) adjectives: acoustic measurements and perception data, *Phonetica*, 44, 1-12.

Heuven, V.J. van (1988) De waarneming van spraak, in M.P.R. van den Broecke (ed.) *Ter sprake, spraak als betekenisvol geluid in 36 thematische hoofdstukken*, Foris, Dordrecht, 73-103.

Heuven, V.J. van and L. Menert (1996) Why stress position bias? *Journal of the Acoustical Society of America* (accepted).

Heuven, V.J. van (1994) What is the smallest prosodic domain?, in P. Keating (ed): *Papers in Laboratory Phonology III: phonological structure and phonetic form*, Cambridge University Press, London, 76-98.

Jones, D. (1918) *An outline of English phonetics*, Cambridge University Press, Cambridge.

Peterson, G.E. and Lehiste, I. (1959) Vowel amplitude and phonemic stress in American English, *Journal of the Acoustical Society of America*, 31, 428-435.

Lehiste, I. and G.E. Peterson (1961) Some basic considerations in the analysis of intonation, *Journal of the Acoustical Society of America*, 33, 419-425.

Navarro Tomas, T. (1932) *Manual de prononciación española*, Centro de Estudios Históricos.

Nooteboom, S.G. (1972) *Production and perception of vowel duration, a study of durational properties of vowels in Dutch*, doctoral dissertation, Utrecht University.

Rabanales, A. (1960) Hiato y Antihiato en el Español Vulgar de Chile, *Boletín de Filología de la Universidad de Chile*, XII, 197-223.

Sluijter, A.M.C. (1995) *Phonetic correlates of stress and accent*, HIL Dissertation Series No. 15, Leiden.

Sluijter, A.M.C. and V.J. van Heuven (1995) Effects of focus distribution, pitch accent and lexical stress on the temporal organisation of syllables in Dutch, *Phonetica*, 52, 71-89.

Toledo, G.A. (1988) *El ritmo en el español*, Biblioteca Románica Hispánica, Editorial Gredos, Madrid.

Wijk, H.L. van (1961) Los bolivianismos fonéticos en la obra costumbrista de Alfredo Cuillén Pinto, *Boletín de Filología de la Universidad de Chile*, XIII.

Willems, L.F. (1987). Robust formant analysis for speech synthesis, *Proceedings of the European Conference on Speech Technology*, Vol. 2, 250-253.