Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers

Hongyan Wang* and Vincent J. van Heuven
Leiden University Centre for Linguistics
* now at Shenzhen University, Shenzhen, P.R. China

1. Introduction

The vowel systems of (Mandarin) Chinese (e.g. Wiese 1997), Dutch and American English differ considerably, both in the number of vowels in the inventory and in the details of their position within the articulatory vowel space, and possibly also in terms of their durational characteristics. When Dutch and Chinese nationals speak English as a foreign language, their pronunciation of English will deviate from the American native norm for English. As part of a larger research project, we are interested in a precise characterization of Chinese as opposed to Dutch-accented English, and in the question how these non-native varieties of English differ from the native norm. Our description of these three varieties of English will be based on objective rather than impressionistic data (as exemplified by textbooks such as Collins & Mees 1981). Specifically, we used acoustic measurements that are known to have clear correspondences with articulatory properties of vowels.

1.1 Objective measurement of vowel quality

Vowel quality can be quantified by measuring the centre frequencies of the lower resonances in the acoustic signal. The lowest resonance of the vocal tract, called first formant frequency or F1, corresponds closely to the articulatory (and perceptual) dimension of vowel height (high vs. low vowels, or close vs. open vowels). For an average male voice, the F1 values range between 200 Hertz (Hz) for a high vowel /i/ to some 800 Hz for a low vowel /a/. The second formant frequency (or F2) reflects the place of maximal constriction, i.e., the front vs.

Linguistics in the Netherlands 2006, 237–248.
ISSN 0929–7332 / E-ISSN 1569–9919

back dimension, such that the F2 values range from roughly 2400 Hz for front /i/ down to some 600 Hz for back /u/. For female voices the formant frequencies are 10 to 15% higher due to the fact that the resonance cavities in the female vocal tract are smaller (shorter) by 10 to 15% than those of a male speaker.

The relationship between the formant frequencies and the corresponding perceived vowel quality is not linear. For instance, a change in F1 from 200 to 300 Hz brings about a much larger change in perceived vowel quality (height) than a numerically equal change from 700 to 800 Hz. Experimental phoneticians and psycho-physicists have developed an empirical formula that adequately maps the differences in hertz-values onto the perceptual vowel-quality (or timbre) domain, using the so-called Bark transformation.[1]

Probably the best known set of formant measurements was produced for American English by Peterson & Barney (1952) for male and for female speakers separately. These authors used the same stimuli that we used, i.e. vowels embedded in a /h_d/ consonant frame. A similar vowel set was recorded for 50 male and 25 female speakers of Dutch by Pols and co-workers in the seventies (Pols, van de Kamp & Plomp 1973 and van Nierop, Pols & Plomp 1973, respectively). Formant measurements for the vowels of Mandarin (Beijing dialect) have become available only recently (Li, Yu, Chen & Wang 2004). Formant measurements for Chinese-accented English (aiming at the American pronunciation norm) were published by Chen, Robb, Gilbert & Lerman (2001). The authors recorded a subset of the American English vowels (eleven monophthongs) in the same /h_d/ monosyllables that we used ourselves. However, their speakers (20 male and 20 female adults) had been living in the USA for at least two years after having received intensive exposure to spoken English in China in order to qualify for the TOEFL test required to enter a university in the USA. This is clearly a different type of ESL speaker than we target in our study, so that it makes every sense that we should measure the formants in our speaker group separately. No formant data have been published so far for Dutch-accented English vowels.

1.2 The problem of vowel normalization

Unfortunately, formant values measured for the same vowel differ when the tokens are produced by different individuals. The larger the differences between two speakers in shape and size of the cavities in their vocal tracts, the larger the differences in formant values of perceptually identical vowel tokens are. Given that the vocal tracts of women are some 15 percent smaller than those of men, comparing formant values is especially hazardous across speakers of the opposed sex. In the present study we have opted for a straightforward

vowel normalization procedure, first used by Lobanov (1971), which is simply a z-normalization of the F1 and F2 frequencies over the vowel set produced by each individual speaker. In the z-normalization, the F1 and F2 values are transformed to z-scores by subtracting the individual speaker's mean F1 and mean F2 from the raw formant values, and then dividing the difference by the speaker's standard deviation. Z-transformed F1 values less than 0 then correspond to relatively close (high) vowels, values larger than 0 refer to rather open vowels. Similarly, negative z-scores for F2 refer to front vowels, whilst positive F2 z-values represent back vowels. In our case, we applied the Lobanov normalization after first transforming the hertz values to Bark values.

1.3 Vowel duration

The vowels of English and Dutch can be divided into two major groups on the basis of their phonological behaviour, which largely correspond with phonetically short (and lax) versus long (and tense) vowels. Typically, the short/lax and long/tense vowels are in paired oppositions. In English, examples of such pairs are /i: ~ ɪ/ and /u: ~ ʊ/. Vowel durations for American English were published by Peterson & Lehiste (1960). Dutch vowel durations were studied by Nooteboom (1972). No systematic study of vowel duration exists for Mandarin vowels, nor are there systematic data on vowel duration in either Chinese or Dutch-accented English.[2]

Since vowel duration plays a potentially important role in marking the tense ~ lax contrast, next to vowel quality differences, we also measured the vowel duration in the tokens recorded in our dataset. Since some speakers speak faster than others, raw vowel duration cannot be used in the comparison. Rather, durations should be normalized within speakers. Here, too, we used z-normalization so that negative normalized durations refer to relatively short vowel tokens, and positive values represent relatively long vowel durations.

Chinese does not exploit length as a vowel feature at the phonological level. We predict that Chinese speakers of English as a second language (ESL) will distinguish less adequately between the short (lax) and long (tense) vowels of English than Dutch ESL speakers, and certainly less than native speakers of English.

2. Materials

For the present experiment we recorded ten male and ten female speakers for each of three nationalities: Chinese, Dutch and American. All sixty speakers

were students in the Netherlands at the time the recordings were made. Dutch and Chinese speakers had not studied English after secondary school. Speakers did not have, or had in the past, regular contact with English-speaking friends or relatives, nor had they ever lived in an English-speaking country. For a full description of the methods used in the experiment see Wang & van Heuven (2003).

A list of words containing 19 full vowels and diphthongs (so excluding schwa) in identical /hVd/ contexts was recorded: *heed, hid, hayed, head, had, who'd, hood, hoed, hawed, hod, hard, hud, heard, hide, hoyed, how'd, here'd, hoored, haired.* The /h_d/ consonant frame is fully productive in English, allowing all the vowels of English to appear in a word or short phrase.

Speakers were recorded on digital audio tape (DAT) in a sound-insulated recording booth through a Sennheiser MKH–416 microphone. Materials were downsampled (16 KHz, 16 bits), and stored on computer disk.

Our recordings contain tokens of 19 vowel types. Given that our speakers, including the Dutch speakers, without having been instructed to do so, used an American-style pronunciation, with r-coloured (retroflexed) vowels (instead of centring diphthongs), there seemed little point in measuring the vowels that were followed by /r/. Therefore we eliminated the tokens representing *here'd, haired, hard, hoored* and *heard*. Next, we decided not to include any full diphthongs, as these would introduce the complication of having to trace the spectral change over the course of the vowels. This eliminated the types *hide*, *how'd* and *hoyed*. What remained is precisely the set that was also measured in Chen et al. (2001). We finally decided also to eliminate the /ɔ:/ type. It appeared that our speakers (both native and non-native) did not systematically differentiate between this vowel and /ɔ/. Moreover, quite a few of our L2 speakers pronounced /ɔ:/ in *hawed* as /haud/, a pronouncing error induced by the spelling which was not detected at the time of the recording.

3. Results: Vowel quality in Chinese, Dutch and American English

Using the Praat speech processing software (Boersma & Weenink 1996), the beginnings and end points of the target vowels were located in oscillographic and/or spectrographic displays. Formant tracks for the lowest four formants (F1 through F4) were then computed using the Burg LPC algorithm implemented in Praat, and visually checked by superimposing the tracks on a wideband spectrogram. Whenever a mismatch between a track and the formant band in the spectrogram was detected, the model order of the LPC-analysis was changed *ad hoc* until a proper match was obtained between tracks and spectrogram.

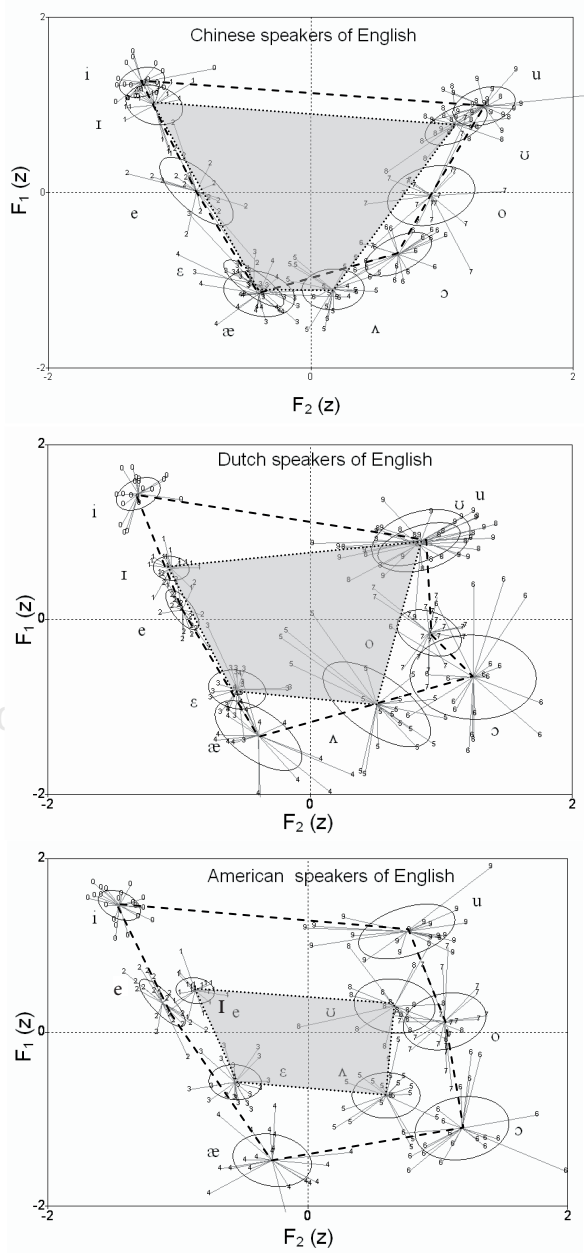

Figure 1. F1-by-F2 plots (Bark-transformed and z-normalized axes) for Chinese, Dutch and American speakers of English. Male and female speakers have been collapsed. Ellipses have been drawn at ±1 SD from the centroids along the two principal component axes. Vowel tokens are linked to the centroids. Dotted lines connect the five lax vowels in each graph (shaded inner polygon), dashed lines connect the tense vowels.

The values for F1 and F2 were extracted at the temporal midpoint of the target vowel, and stored together with the vowel duration for statistical processing.

Formant values were then converted to Bark (see § 1.1), z-normalized within speakers (§ 1.2), and then averaged over the twenty speakers in each speaker group. These mean F1 and F2 values are plotted in acoustical vowel diagrams in Figures 1a–c for Chinese, Dutch and American speakers of English. Each plot contains the position of the ten monophthongs selected as explained in § 1.4.

The vowels of American English are often separated into two length categories, short and long. Phonetically, the four short vowels, /ɪ, ɛ, ʌ, ʊ/ do not only have short durations, they also take up more centralized positions in the vowel space. For this reason, this set may be called 'lax' as well. The other vowels of American English are long and have positions along the outer perimeter of the vowel space. These are, in the present restricted dataset, the vowels /i:, e:, æ, ɔ, o:, u:/. Here, the vowel /ɔ/ is classed as a tense vowel on the grounds that it is a merger of tense /ɔ:/ and lax /ɔ/. Its location in the vowel space (Figure 1c for the American speakers) motivates this choice quite clearly. Following Strange, Bohn, Nishi & Trent (2004), we classified the open front vowel /æ/ as tense. Although it should be phonologically lax (since it cannot occur at the end of a word), there are good phonetic reasons to consider American /æ/ a tense vowel: it is clearly longer than any other lax vowel, and it is also peripheral, that is, on the outer edge of the vowel space. In Figure 1 the six tense vowels have been linked with a dashed line; the lax vowels have been linked with a dotted line.

The **Chinese** ESL speakers' vowels show little spectral distinction between intended /i:/ and /ɪ/. Similarly, there is hardly any spectral difference between intended /ɛ/ and /æ/ nor between /u:/ and /ʊ/.

Moreover, we observe, in Figure 1a, that the tense and lax vowel polygons largely overlap, indicating that the Chinese ESL speakers basically fail to distinguish between the spectrally more peripheral tense set and the spectrally reduced (centralised) lax set.

The ESL tokens produced by the **Dutch** speakers show a clear spectral difference between intended /i:/ and /ɪ/, which is predicted as a case of positive transfer from Dutch to English. There is also a fair degree of separation between intended /ɛ/ and /æ/. Although the separation is not as large as in the native American speech (see below), the success on the part of the Dutch speakers is unexpected. The /ɛ/ ~ /æ/ contrast is typically listed as a learning difficulty (Collins & Mees 1981), and we are surprised to learn that in our ESL speakers some notion of the difference has already been established. Interestingly, the other vowel pair that has traditionally been mentioned as a learning problem, /u:/ ~ /ʊ/, remains completely undifferentiated by the Dutch ESL speakers.

Dutch and English both have tense and lax vowel subsets. Inspection of Figure 1b, however, shows that the tense and lax subsets are not very clearly separated in Dutch ESL. One reason for the relatively poor separation between the subsets is the lack of an /u: ~ ʊ/ contrast in Dutch. The Dutch speakers do not spectrally distinguish between the two, so that here the two subsystems merge (interestingly, the two vowels do differ in their duration — see below). Also, at the lower edge of the vowel space there is little differentiation between more centralized (half) open lax vowels and peripheral open tense vowels as the Dutch ESL speakers do not lower /ɔ/ as much as they should for American English, and at the same time observe insufficient contrast between /ɛ/ and /æ/.

If we now turn to Figure 1c, we notice the **American** native vowels are spectrally much more distinct than those produced by the Dutch speakers, and even more so than the Chinese ESL vowels. There are very large spectral differences between the members of the pairs /i: ~ ɪ/, /ɛ ~ æ/ and /u: ~ ʊ/. Moreover, the figure illustrates quite convincingly that the tense and lax vowel subsets are organised in terms of an outer (peripheral) and an inner (more centralised)

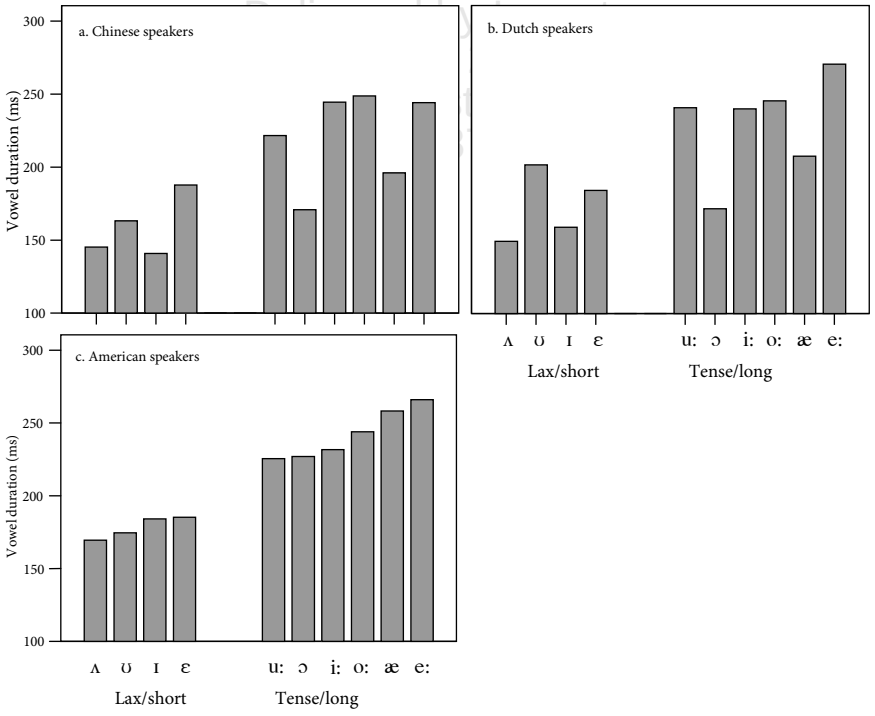

Figure 2. Duration (ms) of four lax and six tense English monophthongal vowels spoken by Chinese (a), Dutch (b) and American (c) speakers of English.

circle. In this respect, too, the L1 speakers clearly differ from both the Dutch and (even more) from the Chinese ESL speakers.

4. Results: Vowel duration in Chinese, Dutch and American English

Figure 2 plots mean duration for each of the ten vowel types, separately for lax and tense categories in for Chinese ESL speakers (panel a), for the Dutch speakers (panel b) and for the American L1 speakers (panel c).

Taking the native speakers as our starting point, Figure 2c clearly shows that the four lax/short vowels have a much shorter duration (with means between 169 and 185 ms) than the six long/tense vowels (with means between 225 and 266 ms). As a result of this, vowels that are spectrally close to each other, such as /e:/ (266 ms) and /ɪ/ (184 ms), are yet acoustically distinct. Note also that when the vowels are ordered from short to long, as has been done in Figure 2c, the increment between adjacent vowels in the figure is never more than 14 ms (which is the difference in mean vowel duration between /o:/ and /æ/). However, the discrepancy between the longest of the short vowels (/ɛ/, 185 ms) and the shortest of the long vowels (/u:/, 225 ms) is 40 ms. These results can be taken in evidence of the phonetic correctness of the subdivision of the American English vowels into the short and long categories made here.

Turning now to the vowel durations produced by the Chinese ESL speakers (Figure 2a), we note that the short vowels are roughly within the duration range of the American L1 speakers. Also, the long vowels are generally within the native range for long vowels, with the exception of the vowels /æ/ and /ɔ/. Interestingly, these are precisely the vowels that distributionally pattern with the short vowels, as they cannot occur at the end of a word in English.

The Dutch ESL vowel durations are rather similar to the Chinese realisations. Again, there are two gross duration categories, one for short vowels with durations less than 200 ms, and one for long vowels with durations in excess of 240 ms. As in the Chinese ESL tokens, the Dutch speakers make the long vowels /æ/ (208 ms) and /ɔ/ (172 ms) too short by American-English standards. Moreover, the Dutch speakers, who did not differentiate between /u:/ and /ʊ/ in spectral terms (see Figure 1b), also have a tendency to make the short /ʊ/ too long (202 ms) — even though this still is still some 40 ms shorter than their mean duration for long /u:/. Unexpectedly, then, it seems as if the Dutch ESL speakers are not more successful in keeping the American-English lax and tense vowels distinct than the Chinese speakers, even though Dutch — unlike Mandarin — is a language with a tense ~ lax subdivision.

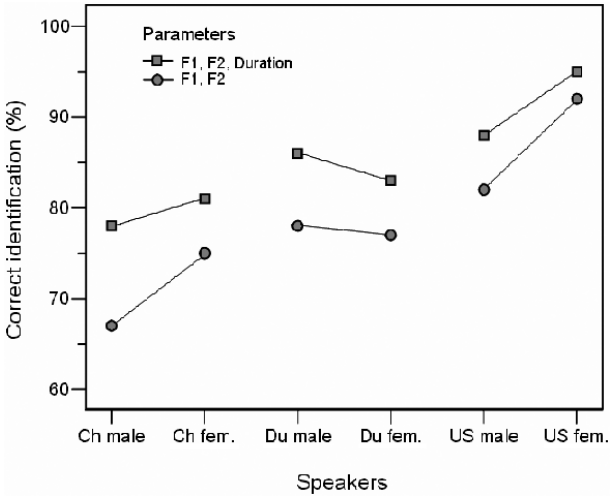

Figure 3. Correctly classified vowel tokens (%) by Linear Discriminant Analysis for Chinese, Dutch and American male and female speakers of English. LDA functions were derived on the basis of F1 and F2 with (squares) and without (circles) duration.

5. Results: Automatic vowel classification

So far we have only been considering the means of the realisations of the vowels — in terms of vowel quality (F1 and F2) and of duration — averaged over groups of twenty speakers. The means do not tell us anything about how well the individual speakers keep the vowels distinct in their pronunciation of English. Figures 1a-c also plot the individual realisation of the vowels in the F1 by F2 plane as scatter clouds, enclosed by spreading ellipses. These were drawn along the principal component axes optimally capturing the directionality of the scatter of the vowel tokens within one vowel type. The ellipses have been plotted at ±1 SD away from the F1–F2 centroids and therefore enclose the most typical 45 per cent (two thirds squared) of the vowel tokens in the category.

The figures show that, generally, the Chinese speakers (Figure 1a) have more overlap between the ellipses of neighbouring vowels than is the case in the Dutch ESL realizations (Figure 1b). The native American L1 speakers have the smallest degree of overlap (Figure 1c), indicating that these speakers keep the ten monophthongs optimally distinct.

We will now attempt to quantify the difference between the three speaker groups in terms of the degree of success in keeping the ten vowels distinct. We have used Linear Discriminant Analysis (LDA, Klecka 1980, Weenink 2006) for this purpose. LDA is an algorithm that computes an optimal set of parameters (called discriminant functions) and automatically classifies objects in

pre-established categories. The more distinct the categories are in the dataset, the fewer the number of classification errors yielded by the algorithm. In the case at hand, the discriminant functions are based on linear combinations of weighted acoustic parameters F1, F2 and duration. Again, before running the LDA, speaker normalization was carried out using the z-transformation on the durations and on the formant frequencies (after Bark conversion).

We ran the LDA algorithm twice. The first time we just included the two spectral parameters as possible predictors of vowel identity, i.e. F1 and F2 (converted to Bark and z-normalized within individual speakers). The second time we extended the set of predictors by also including vowel duration. Figure 3 presents the results of the LDA. The figure shows at a glance that the vowels as produced by the native speakers afford the best automatic identification, those spoken by the Dutch learners can be less successfully identified, and the Chinese ESL tokens are poorest. Adding duration to the set of predictors boosts the correct identification by some 10 percentage points (a little less for the American L1 vowel tokens, possibly due to a ceiling effect). Finally, the vowel tokens produced by the female speakers tend to be more distinct, and therefore better identified, than those spoken by the males. However, there is no such gender effect in the Dutch vowel set.

6. Conclusions

No comprehensive studies are available on the acoustic realisation of English vowels produced by Chinese and Dutch learners, covering both the spectral characteristics of the vowels (in terms of formants) and the duration, and examining the interaction between the two types of parameters in keeping the vowels in the English system distinct. The present study aimed to fill in this gap in our knowledge.

We contrasted learners of English who speak a native language that has a relatively small vowel inventory (Mandarin) and no tense~lax subsets with Dutch learners of English, whose native language has a richer inventory (comparable in size to the English set) and tense versus lax vowels subsets. Chinese and Dutch learners were comparable in the sense that both groups represent non-specialized academic users of English as a foreign language.

Our results shown that the Chinese learners have a rather distorted conception of the American-English vowel system — at least where the ten monophthongs are concerned. The mean positions of the ten vowels in Chinese-accented English are all situated along the outer perimeter of the vowel space, whilst the split of the English system in an outer circle with six tense

vowels and an inner circle with four lax vowels is not observed. As a result, there is insufficient spectral separation between the tense and lax vowel pairs. Interestingly, and unexpectedly, the Chinese learners observe a clear length difference between the four lax vowels of English and the long tense vowels, with the proviso that the two phonologically lax vowels within the tense set, are pronounced short (as would be the case in British English). As a result, the members of the pairs /iː ~ ɪ/ and /uː ~ ʊ/ but not in /ɛ ~ æ/ are acoustically distinct in Chinese-accented English. Consequently, the Chinese vowel tokens are relatively poorly classified by Linear Discriminant Analysis, with roughly 75% correct. This patterning of the results cannot be predicted by a classical type of contrastive analysis of Chinese and English; such an analysis would in fact predict failure of contrast in each of the three pairs.

The Dutch speakers have better acoustical separation of their English vowel tokens, with an average 85% correct classification. Unexpectedly, the Dutch learners observed a reasonably good separation between the members of the /ɛ ~ æ/ pair but predictably failed to keep /uː ~ ʊ/ apart.

The American native speakers have a very clear separation between the tense and lax subsystems, which split is fully supported by a systematic difference in vowel duration. As a result, the American vowel tokens are classified correctly by the LDA in more than 90%.

This acoustical analysis would predict, finally, that Chinese-accented English vowels will be more difficult to identify correctly by human listeners than Dutch-accented English vowels. This prediction is borne out by data of two separate series of experiments which we published in earlier articles (Wang & van Heuven 2003, 2005).

Notes

1. We used the Bark formula as advocated by Traunmüller (1990): Bark = [(26.81 × F) / (1960 + F)] – 0.53, where F represents the measured formant frequency in hertz.

2. However, Elsendoorn (1984) measured vowel durations for six vowel types /iː, ɪ, eː, æ, oː, uː/ in Dutch-accented English spoken by pupils between 12 and 17 years of age at secondary schools.

References

Boersma, Paul & David Weenink. 1996. "Praat, a System for Doing Phonetics by Computer". *Report of the Institute of Phonetic Sciences Amsterdam*, 132.

Chen, Yang, Michael Robb, Harvey Gilbert & Jay Lerman 2001. "Vowel production by Mandarin speakers of English". *Clinical Linguistics & Phonetics* 15.427–440.

Collins, Beverley & Inger Mees. 1981. *The sounds of English and Dutch*. The Hague: Leiden University Press.

Elsendoorn, Ben A.G. 1984. Tolerances of durational properties in British English vowels. PhD diss., Utrecht University.

Klecka, William R. 1980. *Discriminant analysis*. Beverly Hills & London: Sage.

Li, Aijun, Jue Yu, Juanwen Chen & Xia Wang. 2004. "A contrastive study of Standard Chinese and Shanghai-accented Standard Chinese". *From traditional phonology to modern speech processing, Festschrift for professor Wu Zongji's 95th birthday* ed. by Gunnar Fant, Hiroya Fujisaki, Jianfen Cao & Yi Xu, 253–288. Beijing: Foreign Language Teaching and Research Press.

Lobanov, Boris M. 1971. "Classification of Russian vowels spoken by different speakers". *Journal of the Acoustical Society of America* 49.606–608.

Nierop, Dick J.P.J. van, Louis C.W. Pols & Reinier Plomp. 1973. "Frequency analysis of Dutch vowels from 25 female speakers". *Acustica* 29.110–118.

Nooteboom, Sieb G. 1972. Production and perception of vowel duration, a study of durational properties of vowels in Dutch. PhD diss., Utrecht University.

Pols, Louis C.W., Herman R.C. Tromp & Reinier Plomp. 1973. "Frequency analysis of Dutch vowels from 50 male speakers". *Journal of the Acoustical Society of America* 53.1093–1101.

Strange, Winifred, Ocke-Schwen Bohn, Kanae Nishi & Sonja A. Trent. 2004. "Contextual variation in the acoustic and perceptual similarity of North German and American English vowels". *Journal of the Acoustical Society of America* 118.1751–1762.

Traunmüller, Hartmut. 1990. "Analytical expressions for the tonotopic sensory scale". *Journal of the Acoustical Society of America* 88.97–100.

Wang, Hongyan & Vincent J. van Heuven. 2003. "Mutual intelligibility of Chinese, Dutch and American speakers of English". *Linguistics in the Netherlands 2003* ed. by Leonie Cornips & Paula Fikkert, 213–224. Amsterdam & Philadelphia: John Benjamins.

Wang, Hongyan & Vincent J. van Heuven. 2005. "Mutual intelligibility of American, Chinese and Dutch-accented speakers of English". *Proceedings of Interspeech 2005*.1101–1104.

Weenink, David J.M. 2006. Speaker-adaptive vowel identification. PhD diss., University of Amsterdam.

Wiese, Richard. 1997. "Underspecification and the description of Chinese vowels". *Studies in Chinese Phonology* ed. by J. L. Wang & Norval Smith. 219–249. The Hague: Mouton de Gruyter.