DETERMINING NATIVE LANGUAGE BACKGROUND FROM VOWELS AND OBSTRUENTS

Hongyan Wang[1] & Vincent J. van Heuven[2]

[1]Shenzhen University; [2]University of Pannonia, Veszprém Hungary

ABSTRACT

This paper aims to answer the question how acoustic vowel and consonant features can be used to automatically determine the native language (L1) of Mandarin Chinese, Netherlandic Dutch and American speakers of English when very limited materials are available per speaker. Vowel features (intrinsically normalized lowest two resonances and vowel duration) afford 80% correct identification of the speaker's L1 by Linear Discriminant Analysis. Acoustic features of the 16 English obstruents permit 78% correct L1 identification. Correct L1 identification increased to 93% when vowel and obstruent features were combined, making the phonetic feature approach a potentially useful technique for forensic cases with is limited amounts of speech material.

Index Terms— Foreign accent identification; LDA, ALF; acoustic correlates; vowels; obstruents

1. INTRODUCTION

When a foreign language is learned after the age of puberty it is usually the case that the learner's native language interferes with the perception and production of the foreign (or: target) language (e.g. [1, 2, 3, 4, 5]). Typically, the sounds of the foreign language are perceived as exemplars of the sound categories of the learner's native language (e.g. [6, 7]), and sounds of the learner's native language are used as substitutes in the foreign language. The pronunciation of the foreign language is therefore reminiscent of the sounds (and melodies) of the learner's native language, so that the learner's native language can be determined from subtle but systematic deviations in the learner's pronunciation from the norms that apply to the the target language.

In the present paper we aim to study the pronunciation of English by Mandarin-Chinese and by Dutch speakers of English, and compare this with the pronunciation of American native speakers of English. Dutch and English are rather closely related languages within the West Germanic branch of the Indo- European language family, whilst Mandarin, as one of the Sino-Tibetan languages, is genealogically unrelated to, and typologically very different than, either English or Dutch.

Wang [8] studied the production and human perception of the vowels and consonants of English as spoken and perceived by Mandarin, Dutch and American speakers in all nine possible combinations of speaker and listener native language (or: L1) background. Her results showed generally, that the consonants were identified significantly better than the vowels for any combination of speaker and hearer L1. We now ask the question which of the two sets of sounds would provide better cues for the automatic determination of the speaker's native language background. Answering this question may be of interest to several different applications. For one, being able to determine a non-native speaker's language background is often required in forensic applications. Those involved in the international crime scene (including terrorist organizations) often communicate with one another in some form of English – which has evolved into the lingua franca of the 21st century. Knowing the native language of an anonymous caller (e.g. a suspect of a crime) may help law enforcement agencies to narrow down the pool of suspects.

Wang and Van Heuven [9] showed that the L1 background of 20 American, 20 Dutch and 20 Mandarin speakers of English (equally distributed over male and female speakers) could be established at 90 % correct from the lowest two resonances of the vocal tract (F1, F2) and the duration of the ten monophthongal vowels of English. Identification of the speaker's L1 from acoustic properties of the obstruents (lax/voiced and tense/voiceless stops and fricatives) may provide useful, if not essential, additional information on the talker's L1 background. In the present study we have measured a set of 15 relevant acoustic properties of the 16 obstruents of English, and used these to predict the speaker's L1. We try to answer the question whether the speaker's L1 can be more successfully determined from the obstruent properties than from the vowel properties, and whether some combination of vowel and obstruent properties permits even better L1 attribution.

2. METHODS

The data for English spoken with American, Dutch and Mandarin accents were described in detail in [8, 10]. For each language group ten male and ten female speakers were recorded. Non-native speakers were university students who had not specialized in English and had not spent time in an English-speaking environment, i.e. the type of speaker that is the typical ELF user in international settings.

Speakers produced all the 19 full vowels of English in an /hVd/ environment in a fixed carrier sentence Now say h..d again (following [11, 12, 13]). Only the /hVd/ target words were used for acoustic analysis. Each speaker produced one token of each vowel.

Vowel duration and the center frequencies of maximally five formants were extracted; for each vowel token each formant frequency was averaged over the duration of the vowel. Formant frequencies were then psycho-physically scaled in Bark units [14]. In the analysis of the vowel features we concentrated on the ten monophthongs of American English only, i.e. the vowels /i/ (heed), /ɪ/ (hid), /e/ (hayed), /ɛ/ (head), /æ/ (had), /ɒ/ (hod), /ʊ/ (hood), /u/ (who'd), and /ʌ/ (hud).

For each of the 60 speakers we also recorded the 24 onset consonants of English, as spoken in nonsense items /ɑCɑ/ (see [15] for details) embedded in the same carrier phrase Now say /ɑCɑ/ again. From the set of 24 consonants we selected only the 16 obstruents, i.e. the six stop consonants/ p, b, t, d, k, g/, the eight fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ and the affricates /ʧ, ʤ/. The duration (in ms) and intensity (mean and maximum in dB) of the vowel preceding the target C were measured, as was the intensity (mean, max) of the post-consonantal vowel. Duration (ms) and percent voicing were measured for the silent interval of the plosives – as correlates of the tense-lax contrast. Duration, percent voicing, intensity (mean, max) and the spectral mean (Centre of Gravity or CoG) and the spectral standard deviation (SD) were measured (in hertz) on the obstruent noise burst following the silent interval (which was absent in the case of fricatives). The analysis of the spectral composition of the noise bursts was done in order to obtain correlates of the place of articulation (labial, dental, alveolar, pre-palatal) of the obstruents..

3. RESULTS

The analysis and presentation of the results for the vowel features measured for the 60 speakers have been published in [8, 10]. In order to determine the speaker's L1, we entered 30 vowel properties as predictors in a Linear Discriminant Analysis (LDA, [16]). Predictors were the durations of the ten monophthongs (after z-normalization within speakers [17]) and the lowest two resonances of the vocal tract, i.e. F1 and F2, as determined with the Praat speech analysis software [18, 19]. The formants were intrinsically normalized for the individual speaker's vocal tract length (i.e., by dividing both F1 and F2 by F3 – see [20, 21]). Running the LDA in stepwise mode, an optimal model was selected which yielded 80% correct determination of the speaker's L1 using nine predictors: F2(ʊ), F1(ɛ), F1(ɪ), F1(e), D(ʊ), F1(ɒ), D(ɒ), F2(ɛ), and F1(o), in descending order of importance (see Table 1). The spectral properties outweigh the duration cues, even though the latter significantly increased the performance of the model. Spectral cues of lax (short) vowels contribute more to the correct prediction of the speaker's L1 than those of tense (long) vowels. Table 2 presents the confusion matrix of the speakers' predicted L1 (in the columns) against the actual L1 (in the rows), with correct decisions on the diagonal. Figure 1 shows the temporal organization of the English obstruents, for each of the three speaker groups. In the figure, voiceless (tense) and voiced (lax) counterparts have been plotted separately. The Chinese speakers of English stand out by their remarkably long aspiration of the voiceless plosives. The spectral distribution of energy for the noise bursts of the voiceless obstruents can be adequately described in terms of the spectral mean and the spectral standard deviations, i.e. the first two moments of the distribution [15, 22]. No further gain was provided by including also the skewness and the kurtosis.

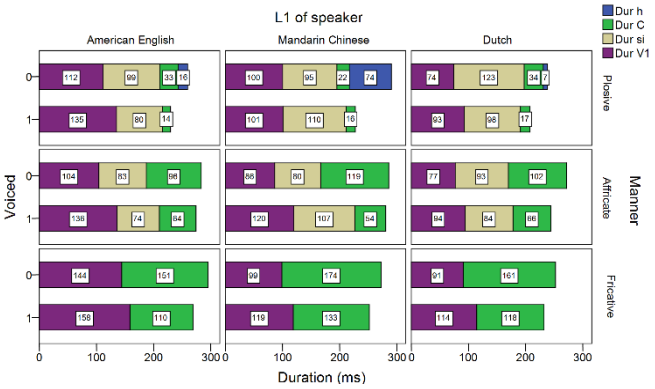

Figure 1. Duration (ms) of pre-consonantal vowel (V1), pre-burst silent interval (si), friction noise (C) and aspiration (h) for American, Chinese and Dutch speakers of English, broken down by manner of articulation and voicing (0 = voiceless, 1 = voiced).

Figure 2 plots the four tense (voiceless) fricatives of English as pronounced by the Chinese, Dutch and American speakers in a two-dimensional plane defined by the spectral mean (CoG, horizontally) against the spectral SD (vertically), for male (upper panels) and female (lower panels) speakers.

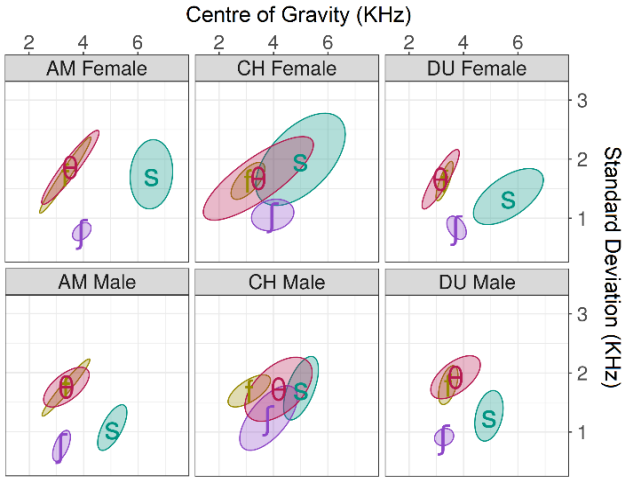

Figure 2. Spectral mean (Center of Gravity, KHz) and spectral Standard Deviation (KHz) for American, Chinese and Dutch speakers of English, broken down by gender of speaker.

The centroid of each fricative is indicated by the location of the phonetic symbol. Dispersion ellipses have been drawn (using [23]) around the centroids at ±1 SD along the first and second principal components of the scatter clouds. It can be

observed that the fricative centroids are further away from one another in the female panels. We therefore decided to apply z-normalization within speakers [19] for the spectral mean and SD for subsequent application in the LDA.

Figure 3A presents the percentage of the duration of the pre-burst silent interval during which glottal pulses could be detected, in separate panels for plosives and affricates (the silent interval is absent in the case of fricatives) for voiced and voiceless obstruents, broken down by the L1 of the speaker group. Figure 3B presents the same information for the percentage of voicing during the obstruent noise burst.

Dutch speakers of English have more pre-voicing in their voiced stops and affricates than American native speakers, whereas Chinese speakers do not differentiate voiced from voiceless silent intervals at all. Chinese speakers have clearly more voicing during the noise bursts of their voiced affricates and (especially) fricatives than the other speaker groups.

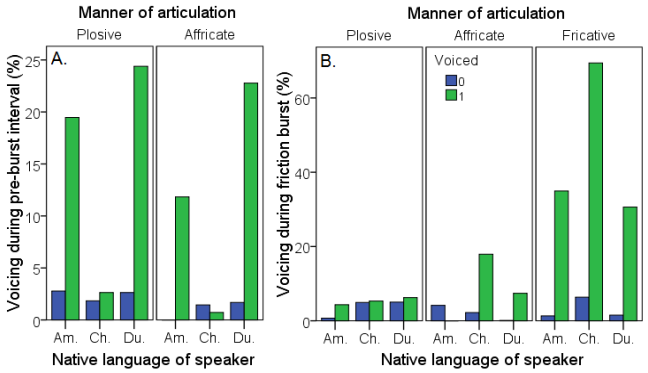

Figure 3. Percentage of voicing during pre-burst silent interval (A) and during friction noise (B) broken down by manner of articulation and voicing of obstruent, for American, Chinese and Dutch speakers of English.

Figure 4 plots the peak intensity of the noise bursts (in dB relative to the loudest abutting vowel), showing that the relative peak intensity of voiceless obstruents is larger than that of the voiced counterparts. The intensity of the noise bursts of the voiced fricatives and (especially) affricates produced by Chinese speakers is weaker than of those produced by the other two speaker groups. This may be related to the circumstance that Chinese has no voiced fricatives and affricates.

Prediction of the speaker's L1 from obstruent features was done by LDA in stepwise mode with six parameters for each of 16 obstruents, i.e. a set of 96 predictors. Only six parameters made a significant independent contribution to the classification of L1. These are shown in Table 1 under 'Obstruent features', in descending order of importance.

The percentage of voicing of the fricative /v/ and of the silent interval of /ʤ/ are the first two predictors, followed by the intensity maximum (relative to the loudest of the two surrounding vowels) of /p/, /ʤ/ and /z/, the latter two separated by duration ratio of the pre consonantal vowel and the total duration of /t/ (silent interval + burst + aspiration).

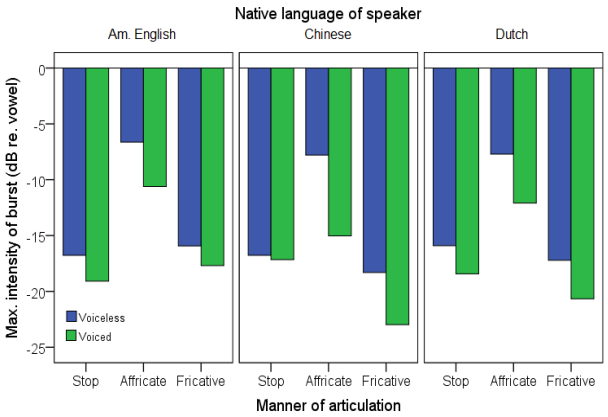

Figure 4. Intensity maximum (re. loudest abutting vowel, in dB) for American, Chinese and Dutch speakers of English, broken down by manner of articulation and voicing.

Table 1. Order of inclusion of acoustic parameters in three LDAs (see text for explanation).

Vowel features		Obstruent features		Combined features	
#	Parameter	#	Parameter	#	Parameter
1.	F2(ʊ)			1.	F2(ʊ)
		1.	Voi%(v)		
		2.	Voi_si%(ʤ)		
		3.	Imax%(p)		
		4.	Imax%(ʤ)		
		5.	VC%(t)	2.	VC%(t)
		6.	Imax%(z)		
2.	F1(ɛ)			3.	F1(ɛ)
3.	F1(ɪ)			4.	F1(ɪ)
4.	F1(e)			5.	F1(e)
5.	D(ʊ)			6.	D(ʊ)
6.	F1(ɐ)				
7.	D(ɐ)			7.	D(ɐ)
8.	F2(ɛ)			8.	F2(ɛ)
9.	F1(o)				

This LDA model affords 78% correct classification of L1. Again, the confusion structure shows that the Dutch-accented sounds are closer to the American sounds than the Chinese-accented counterparts are, which is what one would expect given the greater typological and genealogic similarity between the two Germanic languages.

Finally, a stepwise LDA was performed with the nine vowel and the six obstruent parameters selected in the earlier LDAs combined. No significant improvement of the model could be obtained after the inclusion of eight parameters, seven of which were vowel parameter and only one which was an obstruent parameter (see table 1, under 'Combined features'). This obstruent parameter (VC-duration ratio of /t/) was

included in step 2. The inclusion of F2 of /ʊ/ in step 1 rendered the contribution of all other obstruent features useless. The inclusion of the obstruent feature in step 2, in turn, caused the elimination of two of the earlier vowel features, so that just eight predictors remained.

Table 2. Classification (%) by LDA of native language (L1) background of three groups of speakers, using vowel parameters only, obstruent parameters only, and both types combined. Correct classification in bold face. N = 20 speakers per language. Leave-one-out cross-validation was applied.

	L1 of speaker	Predicted L1		
		American	Chinese	Dutch
Vowels only (80% correct)	American	**75**	0	25
	Chinese	0	**90**	10
	Dutch	25	0	**75**
Obstruents only (78% correct)	American	**65**	5	30
	Chinese	0	**100**	00
	Dutch	25	5	**70**
Vowels + obstruents (93% correct)	American	**95**	0	5
	Chinese	0	**95**	5
	Dutch	10	0	**90**

As is shown in figure 1, the temporal organization of the pre-consonantal vowel and the duration of the plosive, and the differences in this temporal organization between the speaker groups, is not a specific property of the stop alveolar stop /t/. The same system will apply to the labial and velar stops, /p/ and /k/, respectively.

The vowel+obstruent model classifies the speakers' native language background at 93% correct. The combined model performs 13 to 15 points better than the separate vowel and obstruent models, without requiring more predictors.

4. CONCLUSION AND DISCUSSION

In this study we asked the question how well the native language background of American, Mandarin Chinese and Netherlandic Dutch speakers of English can be determined from selected features of vowels and consonants, specifically obstruents in the onset position of a syllable. Forensic applications have been developed which perform L1 detection for a number of languages, but these typically require a large amount of speech data both for reference and for test purposes. The present study takes a different approach by trying to isolate a relatively small number of acoustic features of selected speech sounds, which can be found even if the amount of speech materials available is limited. The results showed that L1 identification by Linear Discriminant Analysis (LDA) was roughly equally successful for vowel features (80% correct identification) as for obstruent features (78% correct). Crucially, however, the combination of the vowel and obstruent features selected by the earlier LDAs increased the percentage of correct L1 identification to 93. This performance compares favorably with that of earlier systems (e.g. [24, 25, 26].

We have tested the feasibility of our approach on English vowels produced by speaker of two languages, Dutch and Mandarin, which are typologically very different from each other (see introduction). It would seem realistic to assume that the set of non-native languages (and thereby the number of different foreign accents in English) can be substantially extended. Given that every language has its own special arrangement of vowels and consonants in the articulatory space, there will generally be a unique combination of acoustic properties that distinguishes the native language from its competitors.

The method may fail for non-native speakers of English who have developed an exceptionally good approximation to the native English norm but this will only be the case for speakers who have learnt English during childhood (so-called early bilinguals, see [27] and references therein) or who were trained to mimic the native pronunciation (e.g. [28]). However, the method should work for the typical educated foreign speaker of English as a lingua franca found in international conferences and business meetings.

ACKNOWLEDGMENTS

This study was sponsored by the National Social Science Fund of China (17FYY009) and the Chinese Ministry of Education Humanities and Social Sciences Planning Project (14YJA740036). We also acknowledge financial support under project TÁMOP 4.2.1.D-15/1KONV-2015-0006 "Development of the innovation center in Kőszeg in the frame of the educational and research network at the University of Pannonia" (subsidized by the EU and Hungary and co-financed by the European Social Fund).

REFERENCES

[1] E. D. Polivanov, "The subjective nature of the perceptions of language sounds", The Hague: Mouton, 223–237, 1974.

[2] R. Lado, *Linguistics across cultures*, Ann Arbor: University of Michigan Press, 1957.

[3] E. Lenneberg, *Biological foundations of language*, New York: Wiley, 1967.

[4] J. Flege, "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," *Journal of Phonetics*, 15: 1547–1565, 1987.

[5] J. E. Flege, "Second language speech learning: theory, findings, and problems", W. Strange (ed.), *Speech perception and linguistic experience: issues in cross-language research*, Timonium MD: York Press, 233–277, 1995.

[6] C. T. Best, "A direct realist perspective on cross-language speech perception," W. Strange (ed.), *Speech perception and linguistic experience: issues in cross-language research*, Timonium MD: York Press, 171–204, 1995.

[7] P. K. Kuhl, and P. Iverson, "Linguistic experience and the perceptual magnet effect," W. Strange (ed.), *Speech perception and linguistic experience: Issues in cross-language research*, Timonium MD: York Press, 121–154, 1995.

[8] H. Wang, *English as a lingua franca: Mutual intelligibility of Chinese, Dutch and American speakers of English* (LOT dissertation series 147), LOT, Utrecht, 2007.

[9] H. Wang, and V. J. van Heuven, "Relative contribution of vowel quality and duration to native language identification in foreign-accented English," *Proceedings of the Second International Conference on Cryptography, Security and Privacy*, Guiyang (in press), 2018.

[10] H. Wang, and V. J. Heuven, "Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers," J. M. van de Weijer, and B. Los (eds.), *Linguistics in the Netherlands 2006*, Amsterdam: John Benjamins, 237–248, 2006.

[11] G. E. Peterson., and H. L. Barney, " Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, 24: 175–184, 1952.

[12] J. Hillenbrand, L. Getty, M. Clark, and L. Wheeler, "Acoustic characteristics of American English vowels," *Journal of the Acoustical Society of America*, 97: 3099–3111, 1995.

[13] Y. Chen, M. Robb, H. Gilbert, and J. Lerman, "Vowel production by Mandarin speakers of English," *Clinical Linguistics & Phonetics*, 15: 427–440, 2001.

[14] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *Journal of the Acoustical Society of America*, 88: 97–100, 1990.

[15] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives," *Journal of the Acoustical Society of America*, 125: 3962–3973, 2009.

[16] W. R. Klecka, *Discriminant Analysis.* Beverly Hills, CA, London: Sage, 1980.

[17] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *Journal of the Acoustical Society of America*, 49: 606–608, 1971.

[18] P. Boersma, and D. Weenink, "Praat, A system for doing phonetics by computer," *Report of the Institute of Phonetic Sciences Amsterdam*, 132, 1996.

[19] P. Boersma,, and V. J. van Heuven, "Speak and unSpeak with Praat," *Glot International*, 5: 341–347, 2001.

[20] G. E. Peterson, "The phonetic value of vowels," *Language*, 27(4): 541–553, 1951.

[21] P. J. Monahan, and W. J. Idsardi, "Auditory sensitivity to formant ratios: Toward an account of vowel normalization," *Language and Cognitive Processing*, 25: 808–839, 2010.

[22] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *Journal of the Acoustical Society of America*, 108: 1252–1263, 2000.

[23] W. Heeringa, and H. van de Velde, "Visible vowels: A tool for the visualization of vowel variation," *Proceedings of Interspeech*, Stockholm, 4034–4035, 2017.

[24] L. M. Arslan, and H. H. Hansen, "Language accent classification in American English," *Speech Communication*, 18: 353–367, 1996.

[25] K. Bartkova, and D. Jouvet, "Automatic detection of foreign accent for automatic speech recognition," *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 2185–2188, 2007.

[26] K. Kumpf, and R. W. King, "Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks," *Proceedings of Eurospeech* '97, Rhodes, 2323–2326, 1997.

[27] V. J. van Heuven, "Perception of English and Dutch checked vowels by early and late bilinguals. Towards a new measure of language dominance," S. E. Pfenninger, and J. Navracsics (eds.), *Future research directions for Applied Linguistics*, Bristol, Buffalo, Toronto: Multilingual Matters, 73–98, 2017.

[28] T. Bongaerts, C. van Summeren, B. Planken, and E. Schils, "Age and ultimate attainment in the pronunciation of a foreign language," *Studies in Second Language Acquisition*, 19: 447–465, 1997.

This paper will appear/appeared as:
Wang, Hongyan & Vincent J. van Heuven (2018). Determining native language background from English vowels and obstruents. *Proceedings of the 2018 IEEE Workshop on Information, Forensics and Security, 11-13 December 2018, Hong Kong*.