Vowel labelling consistency as a measure of familiarity with the phonetic code of a language or dialect

Heuven, V.J.J.P. van; Houten, J.E. van

Vowel labelling consistency as a measure of familiarity with the phonetic code of a language or dialect

Vincent J. van Heuven
Dept. of Linguistics
Phonetics Laboratory
Leyden University
P.O. Box 9515
2300 RA Leiden
The Netherlands

Els van Houten
P.J. Meertens Institute
of Dialectology
Keizersgracht 569-571
1017 DR Amsterdam
The Netherlands

1. Introduction

A considerable part of sociolinguistic and sociophonetic research has been devoted to the study of the use of competing "codes" in multilingual or multidialectal societies. In the typical situation each speaker in the community has access to more than one language or dialect. It is hardly ever the case that the speaker has equal, perfect command of both languages. Very often the speaker acquired one language or dialect as his mother tongue, and only began to learn his second language or dialect at some later stage. The mother tongue may be a regional or urban dialect that has to be replaced by the standard language in particular social settings, or it may be that the speaker has moved to another country and has to learn a second language in order to communicate in his new sociolinguistic environment.

Researchers are often interested in establishing how well such a speaker commands each of his several languages or dialects. As a case in point consider a recent sociolinguistic project carried out in The Netherlands on the diminishing influence of the local Frisian language, which is gradually being replaced by Dutch. Although virtually all Frisians speak both Frisian and Dutch, it was hypothesized that the command of Frisian would be better for older speakers and poorer for younger speakers, both in an absolute sense, and relative to the speaker's command of Dutch. Familiarity with Frisian was then measured in terms of phonetic indices, i.e., by counting the incidence of specific sounds or phonological rules in the speech samples produced by each speaker in the survey.

There are several drawbacks to the use of speech production data as a method of establishing an individual's linguistic command. The method is extremely time–consuming, and relies on auditory judgment by panels of phonetically trained listeners. If the phonetic indices are to be measured in the acoustic domain, the necessary time investment will only increase. And even if the phonetic indices can be measured reliably, it will still be unclear why the speaker's production is deficient. Is it because he cannot articulate the sounds properly due to a purely motoric inability, or is it because he has a poor mental conception of what the speech should sound like? If the true cause of deviant speech production is in the speaker's

poor perceptual representation of the sound system, it would make sense to get at the perceptual representation directly, rather than through his speech production.

2. The perceptual labelling method

One attractive alternative to the use of phonetic production indices is what we have come to call the vowel labelling task. The method was first used in the early 60's by Cohen, Slis & 't Hart (1963) and Delattre (1965). Listeners controlled a vowel synthesizer, and were instructed to manipulate the vowel quality (and duration) until they were satisfied that they had generated the best possible approximation to a particular vowel in their language. On the basis of the results one can map out a vowel space for each listener, and compare individual differences. However, the procedure is extremely tedious since the optimal quality (and duration) for each target vowel has to be found by trial and error. Especially when the language has a large vowel system, the task is too demanding.

Therefore an alternative has been developed (e.g. Blom & Uys, 1966; Schouten, 1975; Ainsworth, 1976; Hombert, 1979), in which the subject listens to a randomized series of vowel sounds which have been synthesized by the researcher according to some methodical plan. Typically vowel duration is held constant at some convenient mean value, while the vowel quality space is sampled along two dimensions, viz. the lowest resonance of the vocal tract (F1, corresponding to vowel height) and the second–lowest resonance (F2, corresponding roughly to vowel backness). The listener's task is to indicate for each of the synthesized vowel sounds, with forced choice, which vowel in the inventory of his language the stimulus resembles most.

We adopted this paradigm in some of our own research, but added a systematic analysis of the consistency with which the listeners labelled repeated tokens of the same stimulus types. The aim of the present paper is to examine more systematically than we have done so far, to what extent the collected labelling consistency data can be used to express an individual's familarity with (the phonetic code) of a language or dialect.

3. Familarity and labelling behavior

Native listeners of a language are eminently able to categorize any (vowel) sound in terms of the vowel inventory of their language. A very popular technique to study the location of the boundary between two particular phonetic categories was developed at the Haskins Laboratories some 35 years ago. In its simplest form the researcher created a continuum between two sounds by changing the value of just one synthesis parameter in small steps. Consider, for instance, figure 1, where we see the results for a manner of articulation continuum going from affricate "chop" to fricative "shop". The 8 stimulus types that made up the continuum were all identical but for one parameter, viz. the duration of the friction portion at the word onset. On

the affricate extreme of the continuum the friction portion was given a duration of 40 ms, while the fricative extreme had a friction noise of 180 ms. The continuum between these extremes was sampled in steps of 20 ms. Several tokens of each of the 8 stimulus types were presented in random succession. Native American listeners were asked to decide for each token whether they perceived it as ''shop'' or ''chop''. The results are plotted in figure 1.

We observe that the extremes are unanimously judged to be instances of either ''chop'' or ''shop'', but that the decision is ambiguous for stimuli in the middle of the continuum. In figure 1 the listeners are undecided for only one stimulus; as a consequence there is a rather abrupt cross–over from fricative to affricate with a very sharp boundary between the two categories along the noise duration dimension. The sharpness of the boundary can be expressed in units along the stimulus axis (here milliseconds), most often in terms of the standard deviation of the cumulative normal distribution that can be fitted to the data points.

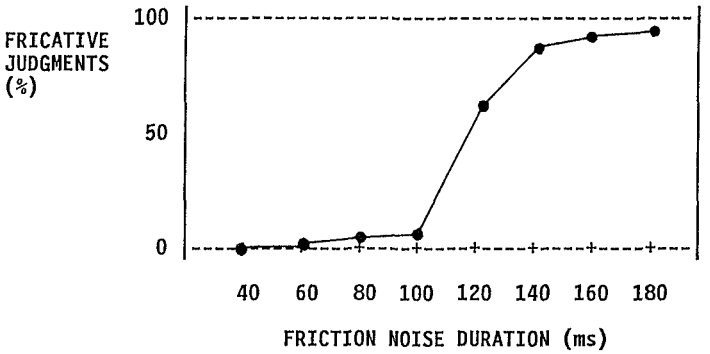

Figure 1, Percent fricative judgments as a function of friction noise duration for a chop–shop continuum. Noise rise time is held constant at 40 ms (adapted from Gerstman, 1957; van Heuven, 1979)

It appears that native and foreign listeners perform this type of task equally well for the end points of the continuum, i.e., the extreme stimulus types. However, it is characteristic of the performance of a non–native subject that the delineation of the phonetic categories is poor. Consequently there is a large area of uncertainty in between categories, and the psychometric functions as in figure 1 have large standard deviations.

This method is practical only for small inventories of phonetic categories, and when the stimulus continua are one–dimensional. When investigating complete vowel systems, the stimulus space is necessarily multi–dimensional, with a large number of categories within the space. It is then virtually impossible to examine boundaries between the multiple categories. The number of responses necessary to trace out all the category boundaries would have to be astronomical. Thus, Schouten (1975), who was able to map out the perceptual vowel space for native and foreign

speakers of English and Dutch (in all 4 combinations), could not find any clear difference between native and foreign subjects in terms of well–definedness of the category boundaries.

It occurred to us that the effect of ill–defined category boundaries would have to come to light as well, or even better, if we measure the listener's response consistency across the entire stimulus set. The wider the margin of uncertainty between two categories, the more often an individual subject will give conflicting responses to repetitions of the same stimulus type.

A first indication of the power of this type of consistency index as a measure of familiarity with a phonetic code was found in van Zanten & van Heuven (1984), where we investigated the perceptual representation of the Standard Indonesian vowel system for groups of speakers from three regional variants of Indonesian. We shall not recapitulate these results, but instead present vowel labelling consistency data collected in two further experiments. We shall see to what extent a simple consistency index can serve to discriminate between groups of subjects that potentially differ in degree of familiarity with a target language. The first experiment examines data that have not yet appeared elsewhere; the second experiment presents consistency data that have already been published, but are treated in rather more detail here than has been done before.

4. Native versus foreign language vowel labelling

A set of 2 x 74 isolated monophthongal vowel sounds was synthesized using a Fonema OVE 1d vowel synthesizer, sampling the F1–F2 plane along both dimensions in 10 steps of 15% increments (F1 ranging between 250 and 879 Hz, and F2 between 700 and 2463 Hz). F1 had to remain 300 Hz below F2, which condition excluded 26 of the 100 logically possible combinations. F3 was set at 600 Hz above F2 with a maximum of 2600 Hz. F4 and F5 were fixed at 3500 and 4500 Hz, respectively. Bandwidths cannot be varied on this type of synthesizer. For each vowel fundamental frequency fell linearly from 200 to 100 Hz over the course of its duration. Each vowel was given two different amplitude envelopes: the vowel onset always contained a linear rise of 33 ms, followed by a steady state vowel portion of 100 ms. However, the offset portion had a fall–time of either 100 ms (short vowels) or 200 ms (long vowels). The series of 148 vowel sounds (preceded by 32 practice items) was recorded on audio tape in two different random orders. Three groups of subjects were asked to indicate for each of the 288 stimulus tokens, which R.P.–English vowel it resembled most, with forced choice from among the 11 English monophthongs. One group of subjects comprised 7 native speakers of R.P.–English currently living in The Netherlands. A second group contained 7 native speakers of Dutch who spoke R.P.–English as a foreign language at an advanced level of proficiency: each member had obtained at least a first degree in English. The third group contained 9 Dutch speakers of English with no extensive training in English beyond their secondary school education (i.e. 6 years of English training at 2 to 3 hours of tuition weekly).

Here we shall proceed directly to the analysis of response consistency, omitting all information on the actual labelling results. We define a simple consistency index per listener by determining how often both presentations of each stimulus type were identified as the same vowel, out of 148 pairs of identical tokens. Figure 2 plots the consistency index for each of the 23 listeners divided into three groups as defined above.

The results indicate that the consistency index discriminates very well, though not perfectly, between the three proficiency groups. If we discard the results of one advanced foreign speaker of English (marked by ''?'' in figure 2, the first author) on the strength of the argument that this subject's experience with the labelling technique is far more extensive than that of all the other subjects, the separation is quite good indeed.

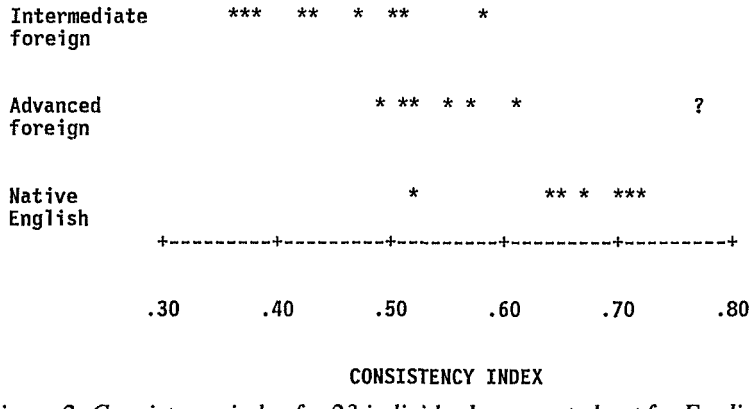

Figure 2: Consistency index for 23 individuals separated out for English native speakers, advanced Dutch learners of English, and intermediate Dutch learners of English.

5. Native versus second language vowel labelling

The second experiment has been described in detail in van Heuven, van Houten & de Vries (1985) and in van Heuven (1985). Here we synthesized a set of 204 vowel stimuli embedded in carrier words, using a Philips MEA8000 speech chip which was controlled by an Apple IIe microcomputer. The F1–F2 vowel quality space was sampled with 34 stimulus points together spanning four vowel continua: front vowels (from [i] to [a]), central vowels (from [y] to [œ]), back vowels (from [u] to [ɑ], and open vowels (from [a] to [ɑ]). F3, F4 and F5 were held constant or varied as a function of F2 as in the previous experiment; formant bandwidths were fixed at mid–range values. Each of the 34 vowels was given 6 different durations between 80 and 200 ms by stretching or compressing the middle portion of the vowel in steps of 24 ms. As before, the 204 stimulus words were recorded on tape in two different

random orders and presented for identification to listeners.

Six native Dutch listeners and 5 Turkish immigrants (who had lived in The Netherlands for at least 8 years) listened to the tape and had to label the vowel in each stimulus word in terms of the Dutch vowel inventory, with forced choice from among the 18 Dutch stressed vowels and diphthongs. The immigrants were all fluent speakers of Dutch, who could read and write in Dutch, and had been selected for their ability to spell the response words without error or difficulty.

The consistency indices came out as indicated in figure 3. Clearly, here the consistency index brings about an excellent separation between first and second language speakers. There is not a single case of overlap between the two groups. Moreover, there is even ample differentiation between individuals within the same listener category.

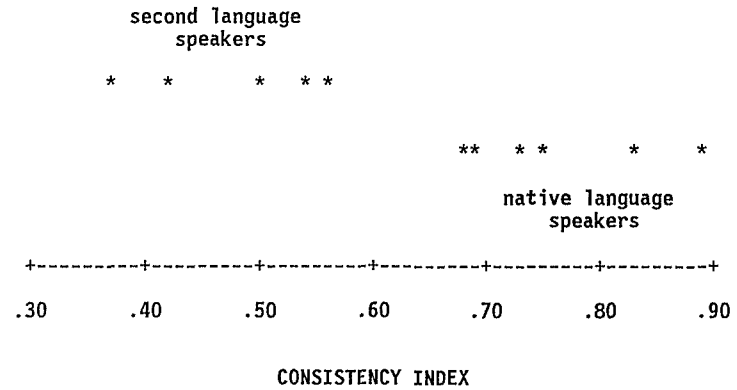

Figure 3, Consistency index for 6 native Dutch listeners and 5 second language speakers of Dutch (see text).

6. Conclusions and discussion

On the strength of the above results we may conclude that a simple consistency index obtained in a vowel labelling task offers a promising tool for discriminating individuals along a scale of familiarity with a phonetic code. The consistency index then captures the well–definedness of an individual's perceptual representation of the ensemble of phonetic categories in a given language.

It remains to be shown that the index also differentiates between the more and less familiar language(s) or dialect(s) spoken by one individual. Unfortunately, there are considerable practical problems in carrying out the necessary tests. Ideally, the same set of synthesized vowel sounds should be labelled by the same individual, once in terms of his native language vowel inventory, and a second time in terms of foreign vowels. The problem then arises how to synthesize a vowel space such that the vowel contrasts of two possibly widely different languages can be

accommodated without ending up with an unmanageably large stimulus set.

There seem to be possibilities to enhance the discriminatory power of the test. One obvious improvement would be to increase the efficiency of the test by leaving out those stimulus points that do not discriminate between e.g., first and second language listeners. As stated above, even foreign listeners soon have an adequate idea of what the end–points of a contrast should sound like; the main problem is always in the representation of the boundary between the categories making up the contrast. Therefore an efficient test would concentrate on stimulus points close to the category boundaries.

A comparison of the results obtained for the two experiments described above seems to indicate that the discriminatory power of the index improves when vowel labelling is performed for vowels embedded in words (experiment 2) relative to isolated vowels (experiment 1). Even though the number of response categories (i.e., vowels to chose from) was larger in experiment 2, native speakers perform better there than in experiment 1, whereas the reverse seems to hold for foreign or second language speakers. Because there may well be other reasons for the polarization in the results, e.g., intrinsic differences between the listener groups, further research is called for.

In summary then, our consistency index for vowel labelling tasks offers a promising and potentially powerful tool in dialectology and sociolinguistics. Once an adequate stimulus space has been synthesized, a relatively short listening test (30 minutes at the most) is all that is required to compute the index. The procedure does not involve expert judges, and the evaluation of the data can be done by computer, if necessary. There are, however, quite a few problems that still have to be clarified before the method can be used on a larger scale.

Acknowledgement

Experiment I was run by A. Besançon, M. Bot, L. van Duyn, S. Dwarkasing, M.–J. Sanders, and H. Welsink as part of a seminar on experimental phonetics at Leyden University.

References

Ainsworth, W.A. (1976). *Mechanisms of speech recognition*, Pergamon Press, Oxford.

Blom. J., Uys J.Z. (1966). Some notes on the existence of a 'universal concept' of vowels, *Phonetica*, 15, 65–85.

Cohen, A., Slis, I.H., Hart, J. 't (1963). Perceptual tolerances of isolated Dutch vowels, *Phonetica*, 9, 65–78.

Delattre, P. (1965). *Comparing the phonetic features of English, French, German and Spanish.* Julius Groos Verlag, Heidelberg.

Gerstman, L.J. (1957). Perceptual dimensions for the friction portions of certain speech sounds, Ph.D. dissertation, New York University.

Heuven, V.J. van (1979). The relative contribution of rise time, steady time, and overall duration of noise bursts to the affricate–fricative distinction in English: a reanalysis of old data, in D.H. Klatt, J.J. Wolf (eds.): *ASA-50 Speech communication papers*, The Acoustical Society of America, New York, 407–411.

Heuven, V.J. van, Houten, E. van, Vries, J.W. de (1985). De perceptie van Nederlandse klinkers door Turken [The perception of Dutch vowels by Turks], *Spectator*, 15, 225–238.

Heuven, V.J. van (1985). Some acoustic characteristics and perceptual consequences of foreign accent in Dutch spoken by Turkish immigrant workers, in J. van Oosten, J.F. Snapper (reds.): *Dutch Linguistics at Berkeley, papers presented at the Dutch Linguistics Colloquium held at the University of California, Berkeley on November 9th, 1985,* The Dutch Studies Program, U.C. Berkeley, 67–84.

Hombert, J.–M. (1979). Universals of vowel systems: The case of centralized vowels, in E. Fischer–Jørgensen, N. Thorsen, J. Rischel (eds.): *Proceedings of the Ninth International Congress of Phonetic Sciences*, Vol. II, Kopenhagen, 27–32.

Schouten, M.E.H. (1975). Native–language interference in the perception of second–language vowels, Doctoral dissertation, Utrecht University.

Zanten, E.A. van, Heuven, V.J. van (1984). The Indonesian vowels as pronounced and perceived by Toba Batak, Sundanese and Javanese speakers, *Contributions of the Royal Institute of Anthropology [Bijdragen tot de Taal–, Land– en Volkenkunde]*, 140, 497–521.