Viability of VOT as a Parameter for Speaker Identification: Evidence from Hong Kong[1]

Winnie H.Y. Cheung **Lian-Hee Wee**

Hong Kong Baptist University
winniehtc@gmail.com

Abstract

This research explores VOT as a speaker-specific property within the context of English-Cantonese bilingualism in Hong Kong. Utterances collected from five individuals for /p, t, b, d/ vary over two languages and four emotional states. Results show that VOT means by themselves appear to be generally useless as a speaker-specific property because there is as much inter-speaker variation as there is intra-speaker. However, in this paper, we have been able to show that the profile of VOT shifts of each phoneme across the two languages is speaker-specific. Extrapolating from this, profiles of VOT shifts across other parameters like emotional states and even vowel adjacency would be likewise speaker-specific. Herein lies the viability of VOT for speaker-identification.

Keywords: VOT shifts, speaker identification, bilingual, Hong Kong English, Cantonese.

1. Introduction

Speakers are rarely conscious of the synchronicity (or lack thereof) of vocal fold vibration and the release of plosives phones. This makes Voice Onset Time (VOT) a likely candidate for forensic speaker identification. Thus motivated, we investigate correlations between four variables which are likely to affect VOT values: speakers, moods, languages and the voicing contrast of plosives within each language.[2] The hypothesis is that VOT values should vary significantly across these variables and that there should a systematic correlation between the VOT values and the variables, with particular reference to speaker-specificity if VOT is to have forensic applications. The results of this research show that for bilinguals, the VOT shift of the same plosive phoneme across the languages commanded by that speaker is potentially speaker-specific. As such, the viability of VOT for speaker-identification is stronger for polyglots than for monolinguals, an area of forensic phonetic studies hitherto rarely explored (but see Kilpatrick 2003 for a closely related study on VOT of English and Spanish bilinguals).

Section 2 explains the experimental design and data collection. Section 3 presents our results and provides a brief discussion on what the results indicate. Section 4 explains the implications and limitations of the work, before ending with a conclusion.

[1] This work is supported by FRG/06-07/I-34. The authors are grateful to the informants (names withheld for anonymity) for being such willing subjects. The authors are heavily indebted to Julian H.Y. Kam for her immense contribution as research assistant, *sine qua non*.

[2] The investigation of language variation is motivated by Moosmuller (1997), who argues that phonological backgrounds are prone to individual variation in speaker identification.

2. Design of Experiment

In this research, we consider monosyllabic words involving single plosive onsets which VOTs are measured under variation of (i) speakers, (ii) moods, (iii) languages and (iv) the voicing contrast of plosives within each language. The last of these four is well-documented and require no further elaboration. Tokens are set in a carrier frame for naturalness and for ease of identifying the boundaries of the plosives under study. One-way ANOVA[3] is used to test if the other three variables also have significant effects on mean VOT values.

2.1 Languages Tested

Since Hong Kong is essentially bilingual in English and Cantonese, these two languages are chosen for study. English and Cantonese both make a two-way distinction in voicing. However, Shimizu (1996:13) reports that the VOT boundary value of English plosives is about 20-40ms, which means that below 20ms, the plosive would be certainly perceived as voiced, and above 40ms as voiceless. Lisker & Abramson (1964) [cited in Cho & Ladefoged 1999] present the distinction in Cantonese as 9-34ms for one category and 79-98ms for the other, so that the boundary value is 35-78ms, significantly higher than that of English. This means that what is voiced for Cantonese would probably sound very much like unvoiced for English, prompting various linguists to claim that Cantonese makes an aspirate-unaspirate distinction while English makes a voice-voiceless distinction.

Thus, a natural question would be on the VOT for a Cantonese-English bilingual. Since both languages make a two-way distinction, would such a bilingual have a VOT divide that matches English or Cantonese, both or neither?[4] The choice of languages here would serve to address this question. In any case, Hung (2000) reports that speakers of English in Hong Kong have a distinct accent. Specific to plosives, the English spoken in Hong Kong (or Hong Kong English, HKE) appears to be more like Cantonese than English. As will be shown in section 3, this is sometimes the case for some speakers and for some phonemes.

In this paper, the English studied would be HKE.

2.2 Subjects

Five individuals (two males and three females) are used in this study. Their details are provided in the table (1):

(1) List of subjects used in this study

Informants	Sex	Age	Education Level
CY	M	19	University
MW	M	21	University

[3] ANOVA is a hypothesis-testing procedure used to evaluate the mean differences between two or more treatments (Woods et al 1986; Gravetter & Wallnau 2000). ANOVA is useful in this research as we look at the differences in VOT values with respect to speakers, mood and language (i.e. three treatments).

[4] Hung (2000), for example, claims that the English spoken in Hong Kong has a VOT contrast that is fundamentally similar to Cantonese.

AH	F	22	University
JK	F	23	University
WC	F	21	University

Only tertiary educated subjects are chosen here because of their more balanced command in both Cantonese and English (for issues on impact of bilingualism on VOT, see Kilpatrick 2003:Chapter 2). More accurately though, the English spoken in Hong Kong is not phonetically identical with standard varieties such as General American or Received Pronunciation. Nonetheless, because Cantonese has such a dominant presence in Hong Kong, the bilingual subjects exhibit dominant bilinguality (definition in (2)) with a preference for Cantonese. This is very typical of the Hong Kong people. The subjects chosen are thus reasonably representative of Hong Kong.

(2) Dominant Bilinguality (Hamers & Blanc 2000:368)
A state of bilinguality in which competence in one language is superior to the competence in the other; note that dominance is not equally distributed for all domains and functions of language.

2.3 Phoneme and Selection of Words

As mentioned above, the two languages studied (English and Cantonese) make a two-way distinction on the voicing contrasts of plosives. As such, our selection of phonemes include the labials /b/~/p/ and the alveolars /d/~/t/. We have left out the set of velars for want of resources, although our results later will show that our conclusions can be easily extended to cover the velar plosives.

In this research, we set up a list of monosyllabic words involving the plosives /p, t, b, d/ as single onsets across the two languages. For each phoneme, there are six token words to be read three times so that an average value may be obtained. The six tokens are equally divided amongst three cardinal vowels [high, front], [low], and [high, back, round] so that our averages would not be skewed by effects of vowels that immediately follow the plosives. Below, (3) and (4) are our list of test words. The phonetic forms in (3) correspond to pronunciations in Hong Kong, and may not coincide with dictionary entries.

(3) List of English Words

plosives	Following vowel	closed syllable	open syllable
p	i	pig [p ik]	pea [p i]
	o	pot [p ot]	paw [p o]
	a	part [p art]	par [p ar]
t	i	tick [t ik]	tea [t i]
	o	top [t op]	tall [t o]
	a	tart [t art]	tie [t ai]
b	i	big [pit]	bee [pi]
	o	bob [pob]	ball [po]
	a	bark [park]	bar [par]
d	i	dig [tik]	D [ti]
	o	dot [tot]	door [do]
	a	dark [tark]	die [tai]

(4) List of Cantonese Words[5]

plosives	Following vowel	closed syllable	open syllable
p	i	[p^{h}ik]	[p iu]
	o	[p ok]	[p o]
	a	[p ak]	[p a]
t	i	[t ip]	[t iu]
	o	[t ok]	[t o]
	a	[t at]	[t a]
b	i	[pik]	[piu]
	o	[pok]	[po]
	a	[pak]	[pa]
d	i	[tik]	[tiu]
	o	[tok]	[to]
	a	[tat]	[tai]

Because vowels are known to affect VOT values (Shimizu 1996:27, 55, 111), we have tried to factor out their effects by balancing each phoneme with equal number of entries for each of three cardinal vowels: [high, front], [low] and [high, back].

2.4 Carrier frame

The English words and the Cantonese words are put into a carrier frame such that the plosive is preceded by a sonorant. This would allow us to see clearly where the plosive begins. The carrier frames are as given in (5).

(5) a. Carrier frame for English
I am going to say ___ once.
b. Carrier frame for Cantonese
ŋɔ: wui kɔŋ ___ jɐt ts^{h}i.

2.5 Speaker Moods

Because part of our study is to examine the viability of VOT for speaker identification, we decided to factor in emotional states to see if there is any covariation with VOT values that might be speaker-specific. To this end, each speaker is required to utter each word (in their carrier frames) across four different emotional states: neutral, happy, sad and angry. Because different individuals might have different "renditions" of the same emotional state, this part of the experiment is somewhat uncertain. However, it should be quite reasonable to assume that variation of moods would either be manifest in the intensity or the pace of utterance, both which would have a direct impact on VOT.

2.6 Tokens

This experiment involves a total of 2880 tokens are taken (=6 words*4 phonemes*5 speakers*3 readings* 4 moods*2 languages). See above sections for details.

[5] Glosses and tones omitted, since selected words that are natural to Cantonese do not lend themselves easily to translation.

2.7 Data Collection

Subjects were asked to utter the stimuli words in their carrier frames separately in a quiet setting where the utterances are also recorded. The software used is Praat (version 4.4.30, but see Boersma & Weenink 2008 for a more updated version) at a sampling frequency of 22050Hz. Because of the large number of tokens required, recordings are made over a period of no more than two days[6] each with intermittent breaks every 15-20 minutes. Recordings for each language are collected separately to avoid having the subjects having to switch between codes. Within each language, every token word is collected for all emotional states (mood) before moving on to another word. While this means that subjects have to switch emotional states back-and-forth, it helps prevent a building up of intensity of moods. For instance, if a subject is required to constantly keep an angry mood, the mood may feed itself so that the intensity of anger would grow over each repetition.[7]

3. Results and Discussion

This section presents the results obtained from measuring the VOT values of recordings collected under the experiment design given in the previous section. One-way ANOVA is used for calculating F-ratios, all at a significance level of $p < 0.05$, giving us an error of less than 5%.

In (6) below, we present the results for the VOT contrasts that set apart the two-way voicing distinction for Hong Kong English (HKE).

(6) F-ratio and F-dist. of HKE across four moods

	Mean HKE VOT value(s)						Significance
/p, t/	Neutral	Happy	Sad	Angry	F-ratio	F-dist	
C.Y.	0.0843	0.0796	0.0927	0.0801	1.7624		N
M.W.	0.0626	0.0591	0.0708	0.0625	1.4767		N
A.H.	0.0542	0.0435	0.0547	0.0620	3.0165		Y
J.K.	0.0727	0.0644	0.0723	0.0715	0.9752		N
W.C.	0.0635	0.0489	0.0682	0.0446	6.4146		Y
/b, d/						2.8387	
C.Y.	0.0218	0.0217	0.0259	0.0208	3.7030		Y
M.W.	0.0142	0.0175	0.0210	0.0240	5.1858		Y
A.H.	0.0168	0.0127	0.0178	0.0137	1.0237		N
J.K.	0.0097	0.0107	0.0091	0.0049	0.8190		N
W.C.	0.0117	0.0055	0.0113	0.0047	6.2485		Y

Table (6) provides the mean VOT values of each speaker (listed along the first column) across four emotional moods (listed along the top row) when in HKE utterances. The phonemes are divided for the voicing contrast, so that the set of VOT values for the upper half applies to the voiceless phonemes (/p, t/ in this case, having left out /k/ in this research). The lower half of the table is for the VOT values of voiced phonemes (/b, d/

[6] We keep things to within this time frame for fear that a longer stretch of time would bring in other complications such as memory lapses.

[7] A method often used in realistic acting first advocated by Constantin Stanislavski.

in this case). The F-ratio is then calculated (with *df* =3, 44, where *df* refers to degrees of freedom) [8] to see if VOT variation across moods for each individual speaker is significant.[9] The results are summarized in the final column, where one can see that for C.Y., VOT variation is insignificant across moods for both voiceless plosives, but not voiced plosives. For A.H, variation across moods is significant for voiceless plosives, but not for the voiced ones. There is no obvious pattern when the variation is significant for each speaker within or across the two kinds of plosives.

Ideally, if VOT is useful for speaker-identification, then there should be systematically no significant variation for each speaker across moods for at least one set (either /p, t/ or /b, d/) of the phonemes. This is clearly not the case given the results in (6), which undermines the feasibility of VOT as a speaker-specific parameter.

Moving on to Cantonese, table (7) presents the results for the VOT contrasts that set apart the two-way voicing distinction (again, *df* =3, 44).

(7) F-ratio and F-Dist. of Cantonese across four moods

	Mean Cantonese VOT value(s)						Significance
/p, t/	Neutral	Happy	Sad	Angry	F-ratio	F-dist	
C.Y.	0.0757	0.0594	0.0651	0.0583	1.5771		N
M.W.	0.0696	0.0528	0.0711	0.0576	6.0910		Y
A.H.	0.0405	0.0325	0.0412	0.0330	2.1276		N
J.K.	0.0628	0.0744	0.0608	0.0311	4.8038		Y
W.C.	0.0474	0.0320	0.0364	0.0341	2.1927		N
/b, d/						2.8387	
C.Y.	0.0188	0.0183	0.0205	0.0196	0.0728		N
M.W.	0.0209	0.0115	0.0146	0.0113	2.6395		N
A.H.	0.0111	0.0087	0.0155	0.0065	6.3932		Y
J.K.	0.0149	0.0087	0.0147	0.0051	6.7450		Y
W.C.	0.0075	0.0055	0.0095	0.0036	2.3656		N

Table (7) is read the same way as table (6). As can be seen from the results in (7), there is no systematic pattern in when moods make significant variation on the VOT means of each speaker, which corroborates with the results in (6). Again, this undermines the feasibility of VOT as a speaker-specific parameter.

Although (6) and (7) do not appear very supportive, it is nonetheless possible that the applicability of VOT for speaker-identification is confined at the level of the phoneme and could be language specific. The results in (8)[10], allows us to check if this may be so for the case of HKE.

[8] *df* = x,y, where the first value is the degree of freedom between treatments (in this case the moods), while the second value is the degree of freedom within-treatments (in this case the number of utterances that produced the VOT means). On how to calculate *df*, see Woods et al (1986) or Gravetter & Wallnau (2000) for an excellent explanation.

[9] In order to see if moods have impact on each individual speaker, the four emotional states are compared using only one informant each time. By this, one could treat each informant as a separate sample. Hence in this case, an independent measures design is chosen over repeated measures design. One could also notice that since F-ratio is calculated each time using only one speaker. We would therefore have five outcomes (five values of F-ratios). The same thing could be said of the results of table (7).

[10] As we want to see if individuality would impact on the VOT means for a given phoneme, speakers (individual conditions) now become the variable or factor that is being examined. Phonemes /p,t,b,d/

(8) F-ratio and F-Dist. of HKE across five speakers

	Speakers	Mean[11] VOT HKE (s)	F-ratio	F-dist	Significance
/p/	C.Y.	0.0801			
	M.W.	0.0562			
	A.H.	0.0495	16.5218		Y
	J.K.	0.0691			
	W.C.	0.0590			
/t/	C.Y.	0.0883			
	M.W.	0.0713			
	A.H.	0.0578	17.5216		Y
	J.K.	0.0714			
	W.C.	0.0535		2.4500	
/b/	C.Y.	0.0208			
	M.W.	0.0177			
	A.H.	0.0098	12.6636		Y
	J.K.	0.0116			
	W.C.	0.0073			
/d/	C.Y.	0.0243			
	M.W.	0.0206			
	A.H.	0.0208	51.6274		Y
	J.K.	0.0057			
	W.C.	0.0093			

In table (8), we compare the VOT means of each individual across the four chosen phonemes of HKE. Here the results appear to be more encouraging in that there are significant differences (*df* =4, 115, where 4 is *df* based on number of speakers, and 115 is *df* based on number of utterances) for all four phonemes across the speakers as indicated by the "Y" in the last column. However, one has to be cautious of jumping to conclusions here since the comparison is made for all five speakers, not for any given pair of speakers. In other words, the "Y" could be simply the result of having one very deviant speaker while the remaining three have mutually indistinguishable VOT values. To check that for any pair of speakers, there is a significant difference in the VOT values, the Tukey's Honesty Significant Difference (HSD, again see Gravetter & Wallnau 2000 for explanation) post hoc test is needed. If the mean difference of any pair exceeds the critical value (i.e. Tukey's HSD, calculated with the formula in (9)), one can conclude that pair is significantly different.

(9) Tukey's Honesty Significant Difference (HSD)

$$\text{HSD}= q\sqrt{\frac{\text{Variance within treatments}}{\text{number of scores in each treatment}}}$$

, where q is a table value called a Studentized range statistic.

The Tukey's HSD is applied to all possible pairings (which is C_2^5=10) of all five speakers. The results are summarized in table (10).

---

become the control factors. Each time only one phoneme is considered, so there are four F-ratios altogether for comparison. Table (11) is understood the same way as here.

[11] The means are calculated for all utterances across four emotional states.

(10) Pairwise HSD across speakers on HKE /p, t, b, d/

Phoneme	Speakers	M.W.	A.H.	J.K.	W.C.
/p/	C.Y.	✓	✓	×	✓
	M.W.		×	✓	×
	A.H.			✓	×
	J.K.				✓
/t/	C.Y.	✓	✓	✓	✓
	M.W.		✓	×	✓
	A.H.			✓	×
	J.K.				✓
/b/	C.Y.	×	✓	✓	✓
	M.W.		✓	×	✓
	A.H.			×	×
	J.K.				×
/d/	C.Y.	×	×	✓	✓
	M.W.		×	✓	✓
	A.H.			✓	✓
	J.K.				×

In (10), each pairing is shown on the table by matching each row to each column. The corresponding cell is marked ✓ if the pair is significantly different, but × if not. Grey cells indicate that the pair has already been considered. From the random spread of ×, it is clear that not all the differences are significant for any given pair of speakers. VOT as a speaker-specific property is unviable even when confined to the level of the phoneme for a specific language. This claim is further supported by the results in (11) when one moves on to investigate Cantonese (*df*= 4, 115).

(11) F-ratio and F-dist. of Cantonese across five speakers

	Speakers	Mean[12] VOT Cantonese (s)	F-ratio	F-dist	Significance
/p/	C.Y.	0.0591	6.3382	2.4500	Y
	M.W.	0.0579			
	A.H.	0.0386			
	J.K.	0.0680			
	W.C.	0.0424			
/t/	C.Y.	0.0702	42.1360		Y
	M.W.	0.0676			
	A.H.	0.0350			
	J.K.	0.0466			
	W.C.	0.0318			
/b/	C.Y.	0.0152	2.1968		N
	M.W.	0.0130			
	A.H.	0.0114			
	J.K.	0.0122			
	W.C.	0.0084			
/d/	C.Y.	0.0233	17.1836		Y
	M.W.	0.0166			
	A.H.	0.0094			
	J.K.	0.0095			
	W.C.	0.0046			

[12] Like (8) above, the means are based on all utterances across the four emotional states.

Table (11) is read the same way as table (8). From table (10), although VOT variation is significant for voiceless phonemes and /d/ across five speakers, it is not so for /b/, which makes it unnecessary to run the Tukey's HSD.

There remains one last hope for the viability of VOT as a speaker-specific property. Since we are working with bilinguals (recall that Hong Kongers are larger English-Cantonese bilinguals), one possibility is that VOT variation for the same phoneme across the two languages commanded by each speaker is specific to that individual. In other words, for speaker X who commands languages L_a and L_b, the same phoneme /p/ could have VOT variations across the two languages, that is different from some other speaker Y. This possibility is available only to bilingual individuals, but not an option for monolinguals. In (12)[13] below, we make the comparisons of VOT means[14] for each phoneme /p, t, b, d/ across HKE and Cantonese.

(12) F-ratio and F-dist. between HKE and Cantonese

		Mean VOT (s)				
	Speakers	HKE	Cantonese	F-ratio	F-dist.	Significance
/p/	C.Y.	0.0801	0.0591	10.9560		Y
	M.W.	0.0562	0.0579	0.1740		N
	A.H.	0.0495	0.0386	10.8294		Y
	J.K.	0.0691	0.0680	0.0173		N
	W.C.	0.0590	0.0424	16.3281		Y
/t/	C.Y.	0.0883	0.0702	15.9234		Y
	M.W.	0.0713	0.0676	1.2872		N
	A.H.	0.0578	0.0350	24.8831		Y
	J.K.	0.0714	0.0466	31.9538		Y
	W.C.	0.0535	0.0318	24.6337		Y
					4.0400	
/b/	C.Y.	0.0208	0.0152	10.2729		Y
	M.W.	0.0177	0.0130	2.9164		N
	A.H.	0.0098	0.0114	0.8815		N
	J.K.	0.0116	0.0122	0.0455		N
	W.C.	0.0073	0.0084	0.3241		N
/d/	C.Y.	0.0243	0.0233	0.0918		N
	M.W.	0.0206	0.0166	3.3538		N
	A.H.	0.0208	0.0094	39.6752		Y
	J.K.	0.0057	0.0095	4.7846		Y
	W.C.	0.0093	0.0046	0.0003		N

In (12), each phoneme is studied separately and the VOT means for each speaker for each language is tabulated. The F-ratio (*df*= 1, 46, where 1 is the *df* based on number of languages and 46 on utterances) together with the F-distribution show no obvious pattern at first blush. However, a careful study will reveal that the profile of VOT shifts for each plosive phoneme across languages appears to be speaker-specific. Before, we continue, we define VOT shift in (13).

[13] Similar to the calculations of F-ratios across four moods, each speaker is being studied separately.
[14] The means here are taken across all four emotional states.

(13) VOT L-Shift

VOT L-shift is the change in the mean VOT of any given phoneme across two languages.[15]

For current purposes, the actual value of the VOT L-shift is not quite as important as if the shift is significant. That said, consider C.Y. and M.W. for example (see Table 14).

(14) Comparing VOT L-shift of C.Y. and M.W.

		Mean VOT (s)				
		HKE	Cantonese	F-ratio	F-dist.	Significance
C.Y.	/p/	0.0801	0.0591	10.9560	4.0400	Y
	/t/	0.0883	0.0702	15.9234		Y
	/b/	0.0208	0.0152	10.2729		Y
	/d/	0.0243	0.0233	0.0918		N
M.W.	/p/	0.0562	0.0579	0.1740		N
	/t/	0.0206	0.0166	3.3538		N
	/b/	0.0177	0.0130	2.9164		N
	/d/	0.0713	0.0676	1.2872		N

In (14), the final column would inform us if the VOT L-shift is significant. There is no significant VOT L-shift for M.W. for all the phonemes /p, b, d, t/. VOT values across all four phonemes remains fairly stable when he switches between Cantonese and HKE, possibly this means M.W. has only one set of plosives for both languages. However, for C.Y. this is not the case. Except for /d/, changing between Cantonese and HKE affects the VOT significantly. This distinguishes C.Y. from M.W as two different individuals.

With only four phonemes studied here, and two possibilities with regards VOT L-shift (either significant or not), there are 16 ($=2^4$) possible profiles. Thus for a HKE~Cantonese bilingual, with these four phonemes, speakers can be divided into 16 categories depending on the profile of what VOT-L-shifts are significant. This means that there is a 6.25% (1/16) chance that any two HKE~Cantonese bilinguals agree in their VOT L-shift profiles for /p, b, t, d/. In the case of the five speakers studied here, it turns out that all five have different VOT L-shift profiles, tabulated in (15) for ease of comparison.

[15] Thus (13) does not apply to changes in VOT across plosive phonemes in the same language.

(15) Comparing VOT L-shift profiles for all five speakers

		VOT L-shift Significance
C.Y.	/p/	Y
	/t/	Y
	/b/	Y
	/d/	N
M.W.	/p/	N
	/t/	N
	/b/	N
	/d/	N
A.H.	/p/	Y
	/t/	Y
	/b/	N
	/d/	Y
J.K.	/p/	N
	/t/	Y
	/b/	N
	/d/	Y
W.C.	/p/	Y
	/t/	Y
	/b/	N
	/d/	N

Evidently, VOT L-shift profiles are constrained by the number of phonemes and by the number of languages commanded by the bilingual (or rather multi-lingual). This is because the number of VOT L-shift profiles is given by the formula in (16).

(16) Number of VOT L-shift profiles = $2^{P}\ C_2^L$

, where P = number of phonemes; and L = number languages > 1

The formula in (16) factors into account the possibility of multilingualism, where if there are 3 languages commanded by an individual, then it follows that there can be three sets of VOT L-shift profiles: one set between L_A and L_B; between L_B and L_C, and between L_A and L_C.

4. Implication and Limitations

4.1 Increasing viability of VOT L-shift Profile

The results in section 3 imply that the viability of VOT as a parameter for forensic speaker-identification is not entirely useless. At the very least, with respect to bilinguals, it would be useful for narrowing down possibilities when there is a match in VOT L-shift patterns.

In this research we have investigated only four phonemes, and by the formula in (16), the chances of a VOT L-shift. Profile match would be only 6.25%. However, HKE and Cantonese both have six plosive phonemes. Extrapolating from the results of our experiment, the number of VOT L-shift profiles would increase to 64 ($=2^6$), so that there

is only a 1.56% chance[16] of a perfect match on VOT L-shift profiles. Increasingly, Hong Kongers are becoming trilingual (adding Putonghua to their repertoire of languages), which would reduce the chance of match to 0.52%. Though not anywhere as rare are two identical snowflakes, two identical DNA samples or two identical thumbprints, a 0.52% chance of perfect match is itself quite useful, it means the test is 99.48% accurate.

Further, since more and more people are learning foreign languages (Japanese, French, German, Korean, Italian and Spanish are among the favorite foreign languages in Hong Kong), the potential for using VOT L-shift profile increases exponentially.[17]

4.2 Collapsing of phonemes to their voicing categories

There are, however, a number of limitations in the present report. The most obvious one is that the velars have been left out. Also, in studying VOT variation across moods, we have not separated out each phoneme, collapsing /p, t/ into one category and /b, d/ into another. The unpromising results in (6) and (7) could well be due to this. As such, it may well be that, contrary to our claim, VOT remains useful for speaker identification across a range of moods, notably by way of something parallel to VOT L-shift. In this case, what is needed would be shift of VOT values across moods (M-shift). If this turns out to be viable, then it would be even more useful than appealing to VOT L-shift profiles since VOT M-shift would apply to monolinguals as well. When applied to bilinguals, the number would be further multiplied by the number of languages in command for intra-language mood-shift.

(17) VOT M-shift profiles = $2^{P}\ C_{2}^{M}$ L
, where M = number of emotional states > 1.

By (17), a monolingual would have a 0.26% chance of a perfect match when calculate over four emotional states.[18] It would be useful for polyglots too, and would significantly increase reliability, since the total number of profiles would increase to the formula given in (18) where L-shifts would also be applicable.

(18) Number of VOT shift profiles =
VOT L-shift profiles + VOT M-shift profiles

For bilinguals, for example, there would only be 0.12% of a perfect match for bilinguals (assuming a set of six plosives over 4 emotional states).[19] This gives a 99.88% accuracy.

There is but one difficulty on working with emotional states. These states are highly subjective and difficult to discern. In our study, four states (neutral, happy, sad and angry) are chosen, but it is very hard to decide if these can serve as cardinal

[16] Chance of a perfect match is derivable by a simply dividing 1 by the total number of possible combinations, then multiplied by 100%.

[17] This of course assumes that different languages have differing VOT values, which is empirically supported by such works as Shimizu (1996) and Kilpatrick (2003).

[18] Calculated from $2^{6}\ C_{2}^{4}$ *1

[19] Calculated from $2^{6}\ C_{1}^{2}$ + $2^{6}\ C_{2}^{4}$ *2, see (18).

reference states the same way cardinal vowels are used in phonetics or cardinal points are used in a compass.

4.3 Impact of vowels on VOT

It is well-known that the quality of vowels that immediately follow the plosive affects the VOT values (Shimizu 1996:27, 55, 111). In this research, the VOT of each plosive is obtained by averaging the values for plosives immediately preceding a high, front vowel; a low vowel; and a high, back vowel. Thus for any plosive P, we have equal tokens of Pi, Pa and Pu profiles, which are then averaged for the mean VOT of that plosive. This effectively masks the individual effect of vowels on plosives (recall section 2.3). Given that VOT L-shift profiles can be speaker-specific, one can envisage the possibility of VOT V-shift where for each speaker, the impact of vowels on the VOT value of each plosive would be different. Again, since vowel quality differs across languages, the formula for VOT V-shift profiles would look very much like (17), given here as (19).

(19) VOT V-shift profiles = $2^{P}\ C_{2}^{V}\ L$
, where V = number of different vowels > 1.

We have not made calculations in this regard, but if useful, the formula in (19) would have to be revised as (20).

(20) VOT shift profiles =
L-shift profiles+M-shift profiles+V-shift profiles

Evidently, this would make it a lot harder for there to be a perfect match in VOT shift profiles, making the use of VOT a viable and easy tool for forensic speaker identification. The formulae for VOT L-shift, M-shift and V-shift can of course be further fine-tuned. For example, in (16), we have conflated all four emotional states and all vowel differences to obtain the VOT L-shift. This need not be so, which would increase the total number of profiles by a factor of V and M. Likewise, in VOT M-shift, we have conflated the vowel differences and in VOT V-shift we have conflated the emotional state differences.

Alternatively, one can imagine simplifying the formulae for VOT M-shift and V-shift by conflating the language differences so that L=1 for all cases. We are not sure how best to approach this at this point. As in all research that relies on statistics, there are limitations in the sample size, sample range and sample type. In these areas, we invite anyone interested to join us.

5. Conclusion

This research begins with the assumption that since speakers cannot consciously manipulate VOT, it is potentially useful for forensic speaker identification. To this end, data for four plosives / p, t, b, d/ across two languages (Cantonese and Hong Kong

English) are collected from five bilingual individuals. Results show that intra-speaker variation of VOT values often overlap with inter-speaker variation, making any simplistic use of VOT measurements for speaker identification naïve. However, if one looks into profiles of VOT shifts across languages, then one arrives at a speaker-specific property. This would, of course, apply only to people with command over at least two languages. This implies that similar approaches may be used to study VOT shifts across other parameters such as emotional states or vowel environments. If applicable, then tabulating VOT shift profiles would be useful for monolinguals as well.

References

Boersma, Paul & Weenink, David. 2008. Praat: doing phonetics by computer (Version 4.4.30) [Computer program]. http://www.praat.org/ (January 31, 2008.)

Cho, Taehong & Peter Ladefoged. 1999. Universals and variation in VOT: Evidence from 18 Languages. *Journal of Phonetics* 27. 207-229.

Hamers, Josiane F. & Michel H.A. Blanc. 2000. *Bilinguality and Biliingualism*. Cambridge University Press, 2nd edition.

Hung, Tony T. N. 2000. Towards a phonology of Hong Kong English. *World Englishes* 19 (3). 337-356.

Kilpatrick, Cynthia. 2003. *Compromised VOT: Variation in a bilingual community*. MA thesis, University of Texas at El Paso.

Lisker, L. & Abramson, A.S. 1964. A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20. 384-422.

Moosmuller, S. 1997. Phonological Variation in Speaker Identification. *Forensic Linguistics* 4 (1). 29-47.

Shimizu, Katsumasa. 1996. *A cross-language study of voicing contrasts of stop consonants in Asian languages.* Seibido, Japan.

Woods, Anthony., Paul Fletcher & Arthur Hughes. 1986. *Statistics in Language Studies*. Cambridge Univ. Press.