Likelihood Ratio-based Forensic Voice Comparison with Cantonese Short-term Fundamental Frequency Distribution Parameters

ZHENG Ruijuan[1] and Phil Rose[1,2]

[1] School of Humanities and Social Science, Hong Kong University of Science & Technology
[2] School of Language Studies, Australian National University
rjzheng@ust.hk philip.rose@anu.edu.au

Abstract

Motivated by a forensic voice comparison experiment using long-term F0, this paper investigates the potential of short-term F0, describing a Likelihood Ratio-based forensic voice comparison experiment using parameters from Cantonese short-term F0 distributions based on between ca. 0.75 to 1.5 minutes net voiced speech. Non-contemporaneous natural speech samples from 25 adult male Cantonese native speakers are used, and a cross-validated log-likelihood ratio cost of 0.59 is demonstrated, showing that the feature may have some use.

Index Terms: Forensic voice comparison, likelihood ratio, Cantonese, short-term F0.

1. Introduction

This paper describes an experiment investigating, within the Likelihood Ratio framework, how well speech samples from the same speakers can be distinguished from speech samples from different speakers on the basis of their short-term F0. It also looks briefly at the question of the effect of including an abnormal voice in the reference sample.

As pointed out in [1] F0 is a popular parameter in traditional forensic voice comparison (FVC), because of the promising results in early research using mean and standard deviation of its long-term distribution, and also because it is (relatively) easily measurable and there is often lots of it. However, mean and standard deviation long-term F0 are not ideal as discriminatory parameters because they can be affected by factors such as state of health, emotional changes, discourse genre and noisiness of the environment: in short, factors that can have an adverse effect on the between- to within-speaker variance ratio by increasing the latter. The considerable potential of additional parameters from a long term F0 distribution, comprising its mean, standard deviation, skew, kurtosis, modal F0 and modal density, were investigated in [1], prompted by the observation that the shape of speakers' long-term F0 distributions remains "relatively constant over recording sessions". Using non-contemporaneous speech from a large number (201) of Japanese speakers, it was shown that an equal error rate (EER) of ca. 10% could be achieved with a Likelihood Ratio-based approach. However, as was pointed out, this was based on a relatively enormous, and forensically unrealistic, amount of speech for each recording: between ca. ten and 25 minutes. Usually forensic speech samples are much shorter. The effect on the EER of using a smaller amount of speech was thus investigated, showing that, although the discrimination performance degrades with shorter available speech, "it still appears possible to obtain a EER of between ca. 20% and 23% with a relatively small amount – less than 15 seconds – of voiced speech".

This paper tries to apply this method to Cantonese F0 data, but using a smaller, more realistic amount of voiced speech (between about three-quarters and 1 and a half minutes), and using more realistic speech. In addition, it differs in three respects from the experiment with Japanese F0. As pointed out in [2], the EER is not considered the appropriate measure for Likelihood Ratio-based detection systems. Instead we report the log-likelihood ratio cost (C_{llr}), which penalises strongly counter-factual LRs, as the proper measure of accuracy/validity of such a system as the one used here. Secondly, in order to reveal the true performance of the system with a C_{llr}, the raw Likelihood Ratios (LRs), or scores, have to be calibrated. This was not done in the original Japanese experiment, where it is clear from the Tippett plot presented that, although delivering a good discrimination, the raw LRs are very badly calibrated (to be fair, the paper does in fact note that calibration is necessary). We therefore have used calibration in this paper. Finally, the original paper only used a single distribution per recording, whereas we have made the task somewhat more difficult in combining several distributions in each recording to get a more realistic estimate of the within-speaker variation.

1.1. Speakers, corpus, elicitation

For this study, we use data from an apriori homogeneous group of 25 adult male Cantonese native speakers aged between 20 and 30 years old, mostly university students.

In collecting the data, we followed the recommendations for forensic voice comparison databases in [3]. Most importantly, of course, our database contains two non-contemporaneous recordings of each speaker, and these were separated by at least a month. Since the database had to be collected and processed within a short space of time, it contains only natural speech rather than the two different speaking styles described in [3], and thus the experiment is designed to simulate only the informal telephone conversation task, rather than three tasks, with comparison between them.

The speech data in this study were obtained by using a quasi-map task requiring our subjects to answer questions about the Hong Kong Mass Transit Railway (MTR) based on a map they were given. The main idea of such a map task was to elicit natural, but controlled speech where the subject's attention was distracted from what they were saying by having to solve a problem. As part of the MTR map task, subjects were asked to give directions of how they would get from A to B, where A & B are the names of MTR stations separated by a complicated journey involving several changes, for example from Chai Wan to Disneyland. This type of task was expected to elicit a longer answer from the speakers, and it did, especially when they got lost and had to backtrack, and it is their answers that constitute the source of the short-term F0. The first recording contained three separate MTR journeys to

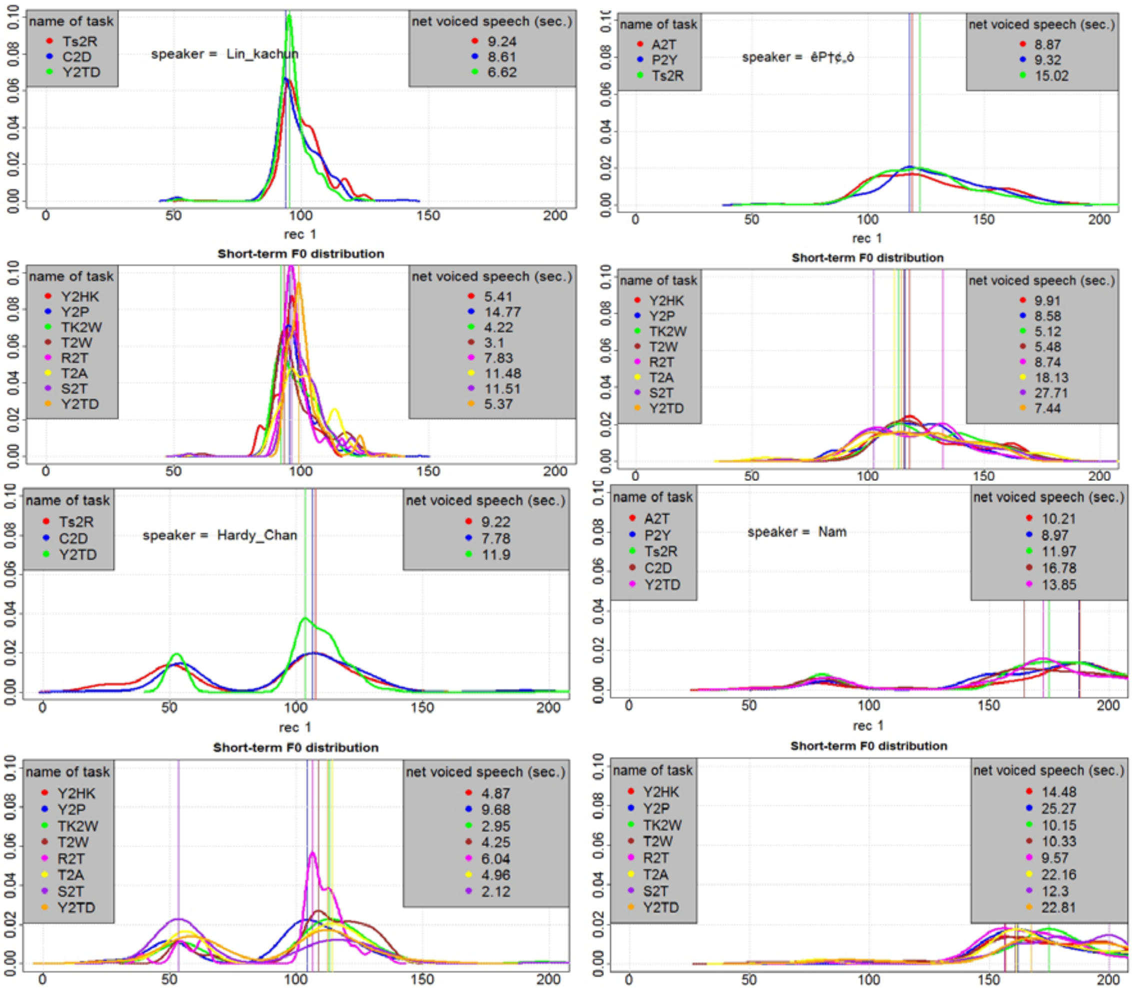

Figure 1: *Non-contemporaneous variation in short-term F0 distributions of four speakers differing in skew, modality and modal F0. Bottom axis is F0, side axis is probability density. See text for explanation.*

describe. In the second, several more were added, in order to elicit more data. Although the recording sessions were separated by the fairly long time of over a month, in order to minimize any learning effect, the journeys to describe differed between the first and second recordings. Most of the participants commented that the questions were more difficult than those of the first round of recording. In a few cases, the speakers failed to give an answer of sufficient duration, or misunderstood the question as one asking how many stations were between A and B. These cases were not included.

The participant speaker was recorded in an acoustically isolated room, while the researcher was located in another room upstairs with an external Edirol UA-25EX digitizing device and the computer. Again following the protocol in [3], the speaker was fitted with a high-quality lapel microphone connected to the digitizing and recording device in the separate room by a long cable. The speaker was then phoned by the researcher on a mobile and they interacted in the map task. In this way, hi-fi recordings (44.1 kHz, 16 bit) of quasi telephone speech responses were obtained and saved as raw PCM .wav files.

1.2. Parametrisation & further processing

The recordings were edited in *Praat*. The stretch of speech corresponding to each separate journey explanation was first identified, and then F0 and corresponding duration extracted at every 0.01 second using the *pitch* (sic) *listing* function. With the exception of one speaker, it was found that a setting of 50 Hz to 300 Hz gave good results for F0 extraction: it excluded most of the incorrect F0 extractions from noise in voiceless segments and is low enough to pick up genuine examples of low F0 from creak. For one speaker with an audible high pitch, it was clear that he had many F0 values above 300 Hz, and the upper limit of his setting had to be changed to 350 Hz. This is the abnormal voice this study wants to pay a little more attention to. The sampled F0 and duration values were then saved in .txt files and manipulated and annotated for further processing in excel files. We tested both raw F0 and $\log_{10}$ F0.

We followed [1], in parametrising the shape of the F0 distribution with six continuous variables: mean, mode and standard deviation F0, skew and kurtosis, and probability density maximum. Likelihood ratios were derived from these variables with the multivariate likelihood ratio described in [4], which was developed to handle the expected dependencies between the variables, and were then calibrated with logistic-regression. A leave-one-out cross-validation was used.

2. Results

2.1. Between & within-speaker variation: duration

Speakers differed in the amount of time they needed to explain

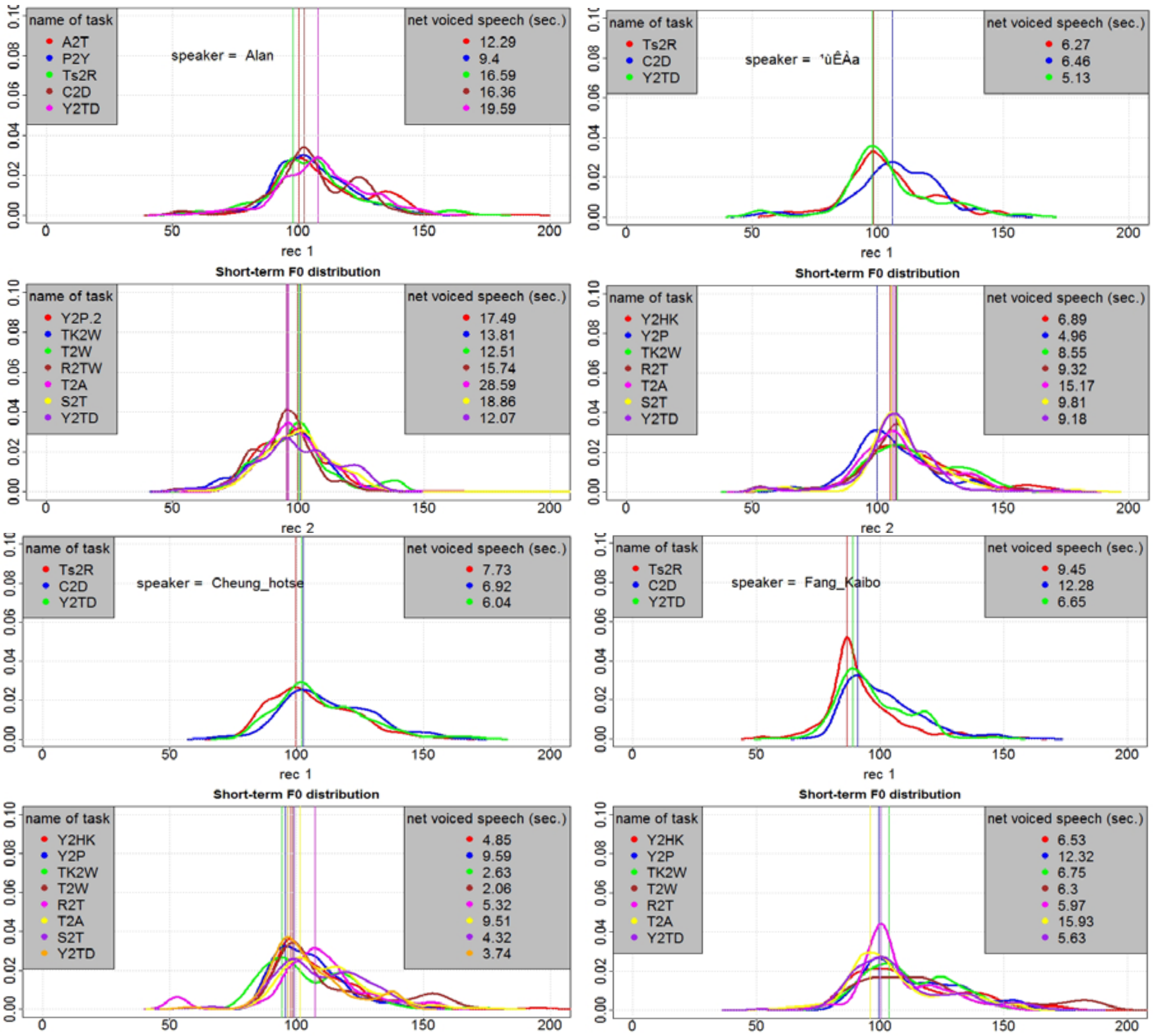

Figure 2: *Four speakers differing little in short-term F0 distributions. See text for explanation.*

how to get from one station to the other. This was partly because some were more verbose than others, and partly because some spoke more quickly; it was also perhaps because some were cognitively more adept at the task. The mean net voiced speech available in the first recording was about ¾ of a minute: 41 sec. The magnitude of its standard deviation (24.3 sec.) suggests a fairly wide range of values – the least amount of speech being ca. 13 sec. and the greatest ca. 109 sec. In the second recording, with more journeys to describe, there was more voiced speech: a mean of about 1.5 minutes (85 seconds), again with a large standard deviation of about 48 sec. Languages differ in the amount of voiced speech they contain. Cantonese has been estimated at the rather low value of 41% [5]. The amount of actual speech necessary for the strength of evidence results we obtained, therefore, would be the net voiced speech values quoted divided by 0.41.

2.2. Between & within-speaker variation: F0

The speakers' short-term F0 distributions showed a certain amount of between-speaker variation in overall shape, with some fairly big differences in skewness and modality, the latter presumably from residual creak. As might be expected from such a homogeneous group, there was not a great amount of variation in mean or modal F0, with most values lying around 100 Hz. The eight panels of Figure 1 show four speakers with rather different distributions. Each speaker is represented by both their non-contemporaneous recordings, with the first on top, and within each non-contemporaneous recording the F0 distribution of each of the journey-tasks is shown separately with a different colour. The tasks are named with an abbreviation for the journey, for example "Ts2R" means how to get from Tsuen Wan to the Racecourse, and the amount of net voiced speech the speaker used in each of their tasks is shown on the top right of each panel. The two speakers in the top two panels differ primarily in kurtosis, the left one being leptokurtic and the right platykurtic. The speaker in the two panels at the bottom left is strongly bimodal, and creaks a lot, which is presumably the origin of their lower mode.

The voice in the bottom right two panels has a very high fundamental, and actually sounds mostly female, although it does have a residue of low F0 values. This voice presents an interesting problem, as it appears to be an abnormal voice occurring in a small set of speakers who were chosen only on the basis of them being easily available young male Cantonese speakers. (We say "appears to be", because we can only at the moment judge abnormality from what they sound like – there is no automatic cut-off in mean F0 values that marks "abnormal" male voices!) Perhaps the 1 in 25 incidence reflects the population; in which case the reference sample should include it. Perhaps it does not, in which case it should be excluded. It is of interest, therefore, to see what the effect is of including the abnormal voice in the reference sample, and so the cross-validated testing was carried out both with and without this voice (i.e. with 25 and 24 speakers).

Of course there were many speakers who did not differ much in their short-term F0 profile. Figure 2 shows the four who differed the least, to the extent that their different-speaker comparison gave the largest counter-factual $\log_{10}$ MVLR of about 0.8 (for the two speakers in the top two panels) and the third largest counter-factual different-speaker LR of about 0.7 (for the two in the bottom two panels).

As can be encouragingly seen from figures 1 and 2, the speakers did not generally show very much variation in the space of a month, with the profiles of their distributions being quite similar. This is similar to the results in [1].

2.3. System performance

The relationship between non-contemporaneous within-speaker variation and between-speaker variation resulted in the MVLR resolution shown in the calibrated Tippett plot in Figure 3. This is for the $\log_{10}$ F0 values, which gave a slightly better C_{llr}. Interestingly, removing the abnormal voice had very little difference, improving the C_{llr} *by* 0.07. Although it is not stellar, the *Cllr* magnitude is easily below unity (indicating that the system is giving some information) and comparable to single vocalic segment C_{llr}*s*. Providing that the expert considers that the genre of the speech data is comparable, as in this task, then short-term F0 is a possible feature to be used in real case-work in conjunction with other features. This is because the LR magnitudes are rather weak. The maximum same-speaker LR does not much exceed $\log_{10}$ 1, and 50% of the different-speaker LRs are also above about $\log_{10}$ -1. One would need unusually advantageous prior odds for these LR ranges to be of benefit on their own, and, although we have a quantification of the accuracy of the feature in its C_{llr}, we also do not yet know its precision [6]. There is presumably enough logarithmicality in the short term F0 to make it sensible to log-transform the values before using them in a real case.

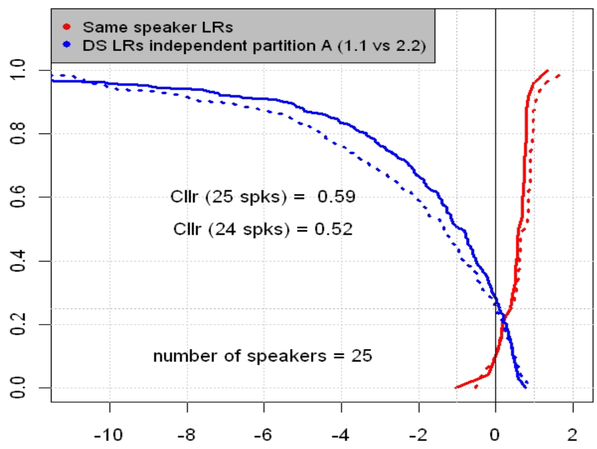

Figure 3: *Tippett plot for non-contemporaneous variation in Cantonese short-term F0 distributions. X axis = Log_{10}LR greater than ...; y axis = cumulative proportion (different-speaker comparisons) or 1- cum. prop. (same-speaker comparisons; dotted line = without abnormal voice.*

3. Summary

A pilot forensic voice comparison experiment investigating short term F0 in Cantonese has been described, and it has been shown that, as hoped, comparisons between non-contemporaneous F0 distribution profiles from the same speaker mostly have LRs greater than unity, and comparisons between F0 distributions from different speakers generally have LRs less than unity. The important metric for this kind of comparison – the C_{llr} – shows that the system is providing useful information, although not a great amount of it.

The main limitations of study are the small number of speakers used, and the differing amounts of speech tested. Both these can be improved with future testing. In particular the way we have structured the data makes it easy to control the amount of speech by including or excluding data from separate journeys. Another limitation is that the recordings were clean. It is not clear how much F0 is affected by the telephone transmission that is normally encountered in forensic work. We would expect it to be minimal with landline, but the algorithms used to encode F0 in mobile phones are a very different thing and almost certainly introduce distortion. With our clean recordings, however, these can now be processed through a mobile system to see what the effects are. We are aware of the crude nature of our parametrisation in that it is based on normality assumptions, and also in that there are going to be strong dependencies between many of the variables. *Eppur*, this paper has shown that such a parametrisation does actually resolve some speaker-specific information. A better way of capturing the shape of distributions like these might be with GMMs, and this should definitely be the next thing we test.

4. Acknowledgements

This paper reaches its readers only with the generous assistance of many nice people. The first author would like to express her sincere gratitude to all her friends who took their time to participate in this project and provide the sound samples: John Lai, Yuan Mai, Kelvin Wong, and William Wu. We especially thank the Hong Kong University of Science and Technology for making it possible for us to run this experiment as part of their postgraduate course *Forensic Voice Comparison in Cantonese*. Thanks also to our referees for their time, some of whose useful suggestions we have incorporated. The paper was written using findings from *Australian Research Council Discovery Grant* No. DP0774115.

5. References

[1] Rose, P., Kinoshita, Y., and Ishihara, S. “Beyond the long-term mean: Exploring the potential of F0 distribution parameters in traditional forensic speaker recognition,” Odyssey Speaker and Language Recognition Workshop, Stellenbosch, 2008.

[2] Morrison, G.S. Forensic voice comparison. In I. Freckelton, & H. Selby [Eds.], Expert Evidence (Ch. 99), Thomson, 2010.

[3] Morrison, G. S., Rose, P., & Zhang C. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Australian Journal of Forensic Sciences, 12: 1-13, 2012.

[4] Aitken, C.G.G. & Lucy, D. “Evaluation of trace evidence in the form of multivariate data”, Appl. Statistics 53(4):109-122, 2004.

[5] Rose, P.J. "How effective are long-term mean and standard deviation as normalisation parameters for tonal fundamental frequency?" Speech Communication, 10, 1991.

[6] Morrison, G.S. “Measuring the Validity and Reliability of forensic likelihood-ratio systems”, Science & Justice 51 (3): 91-98, 2011.