

# Analysis of Type-II Errors for VQ-Distortion Based Speaker Verification

Michael Wagner, Fangxin Chen, Iain Macleod, Bruce Millar, Shuping Ran,  
Andrew Tridgell and Xiaoyuan Zhu

**Abstract**—This paper investigates the distribution of inter-speaker distances for the “training” speakers of the TIMIT speech database. The analysis is based on a speaker verification paradigm with 8 speakers serving as customers and the remaining speakers serving as impostors. The distance measure used is an average variance-weighted vector quantisation (VQ) distortion. It is found that the interspeaker distances correlate significantly with the differences of fundamental frequency (F0) between the speakers. Moreover, the shape of the distribution of impostor distances is largely determined by the customer’s F0. A distinct asymmetry of VQ distances is observed between low-F0 customers and high-F0 impostors on the one hand and high-F0 customers and low-F0 impostors on the other. The type-II error function is estimated from a sample of impostors with similar F0 to the customer. Dialect difference within TIMIT is not found to contribute significantly to VQ distance.

**Keywords**—Speaker verification, Vector quantisation.

## 1. INTRODUCTION

While a number of studies have reported on the performance of speaker recognition systems under a speaker identification paradigm, it is more difficult to evaluate the performance of such systems under a speaker verification paradigm. For a speaker verification system it is necessary to estimate both the type-I (false rejection) and type-II (false acceptance) error rates which depend on the variable distance threshold that separates the acceptable utterances for a given speaker from those that the system will reject. Since the population of “impostors” is essentially open, speaker verification studies based on small groups of speakers often lack reliable estimates of the type-II error.

This study aims at estimating the type-II error function for a speaker verification system assuming that impostors may originate from the entire population of native speakers of American English. Ideally, distances would have to be computed for utterances by every possible impostor against the training data of each of the enrolled “customers” of the verification system — a proposition which is clearly infeasible. In practice, it is essential to be able to estimate the type-II error function for a given speaker from comparisons of that speaker with a small group of impostors drawn from the general population.

Questions therefore arise about how impostor distances are distributed in the general population, whether such distances form clusters for certain subsets of the population, e.g. gender or dialect groups and whether it is possible to select a “close” group of impostors in order to estimate the important low end of the type-II error function.

The authors are with the TRUST (Technology for Robust User-conscious Secure Transactions) Project at the Australian National University, Canberra, ACT 0200. Email: miw@trust.anu.edu.au

TABLE I  
EXPERIMENTAL DATA.

<i>Spkr</i>	<i>Dialect</i>	<i>Sex</i>	<i>F0</i>
dcm	dr7	male	87
bcg	dr8	male	111
rwa	dr3	male	138
bmj	dr4	female	160
rcg	dr1	male	187
pjf	dr2	female	212
lmk	dr5	female	236
sbk	dr6	female	263

This study investigates a customer set drawn from the TIMIT database [1]. The TIMIT data are not often used in speaker recognition or speaker verification experiments as the database does not contain repeated recordings of speakers for adequate characterisation of intra-speaker variance. Because this study investigates the type-II error alone, only inter-speaker variance is of interest, and TIMIT does provide limited material for a large range of speakers.

## 2. METHOD

The speech of each of the 462 “training” speakers in the TIMIT corpus was analysed by firstly computing mel-frequency cepstral coefficients (MFCC), comprising 20 coefficients derived by cosine transform of energy in 110 mel bands spaced 55 mel apart over the range 55 to 2310 mel, for all 10 sentences spoken by that speaker, and secondly, computing the average F0 for the sentence SA1 spoken by all speakers.

Eight reference speakers, or “customers”, comprising 4 females and 4 males, were chosen with each of the eight TIMIT dialect regions being represented. In addition, the speakers were selected such that their average fundamental frequencies represented a cross section of the range of fundamental frequencies found in TIMIT as shown in Table I. A VQ codebook of size 128 was built for each of the reference speakers, using the algorithm documented by Linde et al. [2] on the 20 MFCCs.

The remaining 454 speakers were taken as test speakers, or “impostors”. An average variance-weighted distortion was computed for the sentences SA1 and SA2 of each impostor against the codebook of each of the 8 reference speakers, resulting in a matrix of 8 times 454 inter-speaker distances. The 8 intra-speaker distances were also determined for purposes of comparison.

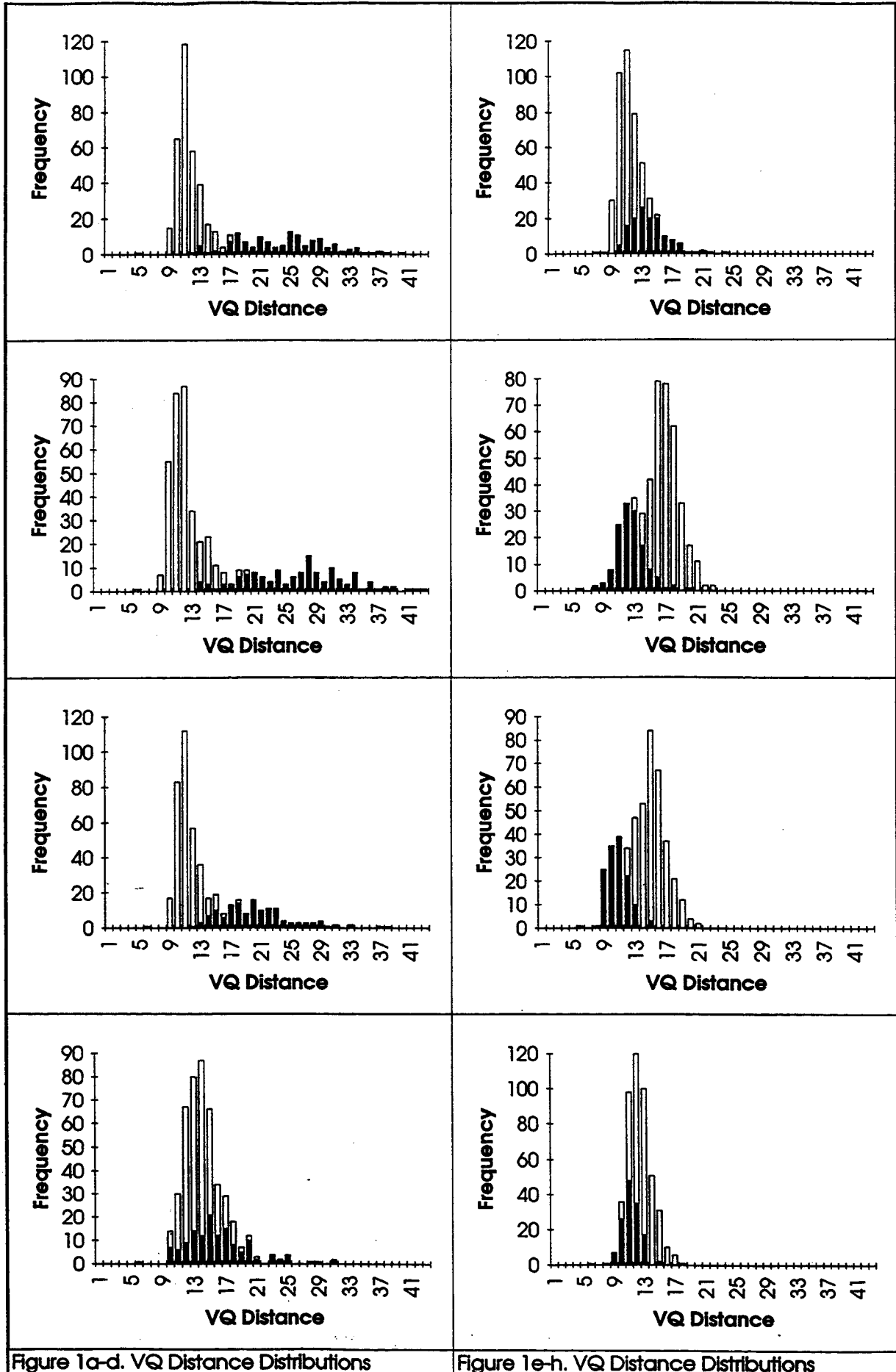


Figure 1a-d. VQ Distance Distributions

Figure 1e-h. VQ Distance Distributions

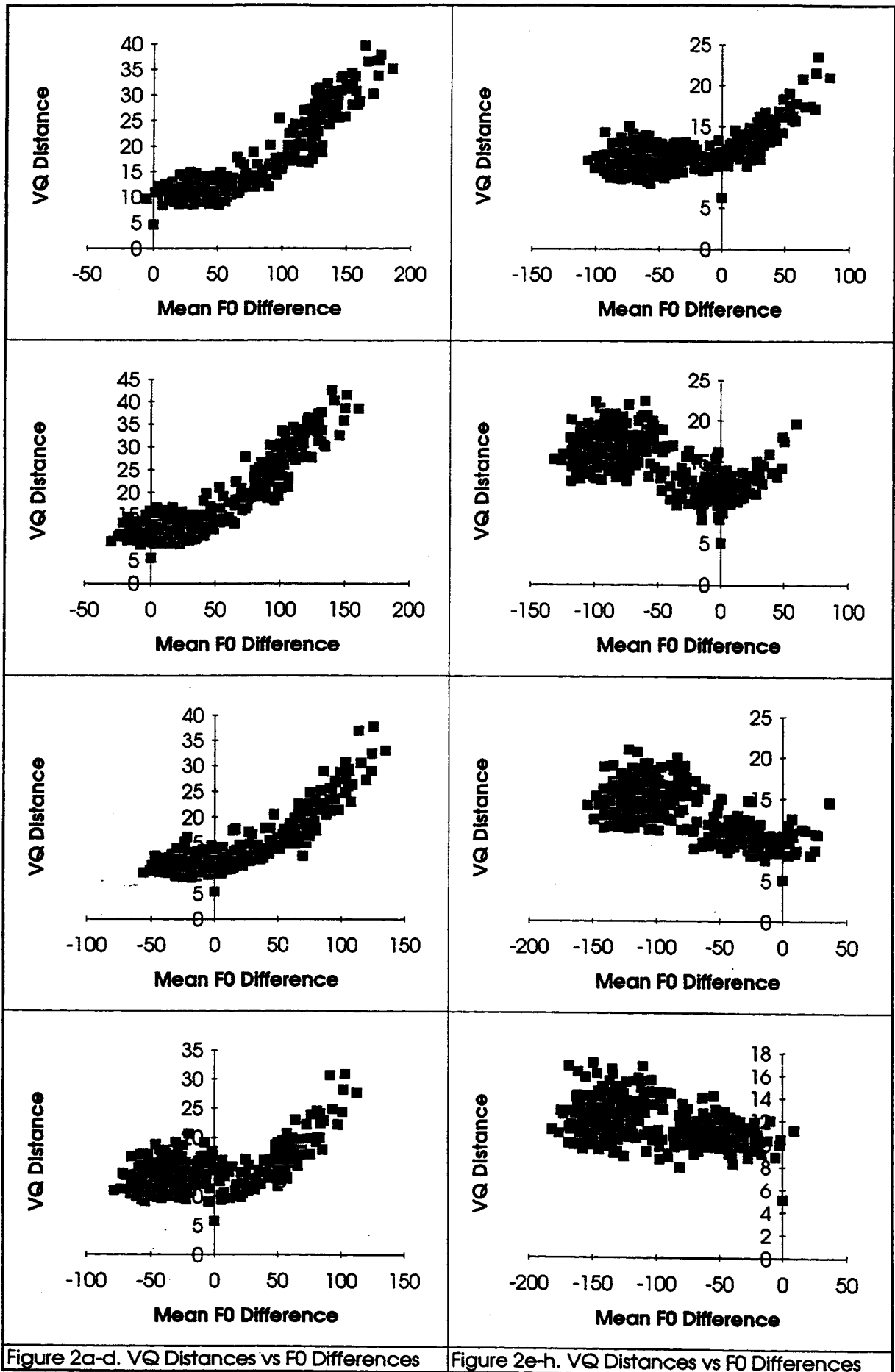


Figure 2a-d. VQ Distances vs F0 Differences

Figure 2e-h. VQ Distances vs F0 Differences

### 3. RESULTS

The distributions of impostor distances for the 8 customers are shown in histogram form in Figure 1. The distributions of male impostor distances are shown by the white bars and the distributions of female impostor distances are shown by the black bars.

It can be seen that male impostors exhibit lower distances to low-pitch customers (Fig. 1a-c) and that female impostors exhibit lower distances to high-pitch customers (Fig. 1f-h). Figure 1 also shows that the distributions of male impostor distances are unimodal and approximately Gaussian for all customers while female impostors have such distributions only for high-pitch customers. For low-pitch customers, female impostors show very wide distance distributions with distances extending to large values. No such large distances are recorded between male impostors and high-pitch customers.

The relation between VQ distance and average F0 was further explored by recording the VQ distances between impostors and customers as they depend on the F0 differences between impostors and customers. The resulting scatter plots for the 8 customers are shown in figure 2.

Figures 2a-e show that VQ distances that are larger than about 20 only occur for customers whose average F0 is more than about 50Hz higher than that of the customer. There is a steep increase of VQ distance for F0 differences between 50Hz and 200Hz. In contrast, the distribution of VQ distances is comparatively flat for negative and small positive F0 differences between impostors and customers. Interestingly, Figures 1e-h show that there is only a small increase of VQ distance for F0 differences of -50Hz to -200Hz between impostors and customers.

These results are shown cumulatively by the scatter plot in Figure 3 which shows all VQ distances against all F0 differences between impostors and customers.

In contrast, dialect differences within TIMIT were not found to contribute significantly to interspeaker differences. The 440 VQ distances between speakers of the same dialect (DR1 to DR7 in TIMIT) were distributed with a mean of 13.43 and standard deviation of 4.19 while the 3080 VQ distances between speakers of different dialects had a mean and standard deviation of 13.77 and 4.73 respectively.

Finally, the type-II error functions were determined for each customer from the distribution of VQ distances. For practical speaker verification systems, only that part of the type-II error curve representing low levels of false acceptance is of interest. The results show that this range of the type-II error curve can be determined by considering a sample of impostors with an average fundamental frequency close to that of the customer.

Figure 4 shows the low error range of the type-II error curve for speaker 2 as it was determined from all 461 impostors. It can be seen from figures 4 and 2b that all impostors within that range have an average fundamental frequency within 35Hz of that of the customer. This range of the error curve can therefore be accurately estimated using a sample of impostors whose average fundamental frequencies are within 35Hz of that of the customer.

Work is currently underway to investigate further the asymmetry of the dependence of VQ distance on the F0 difference between impostor and customer.

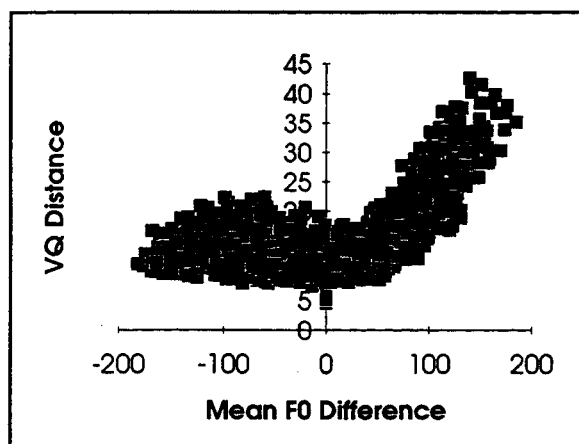


Figure 3. Cumulative VQ distances vs F0 diff.

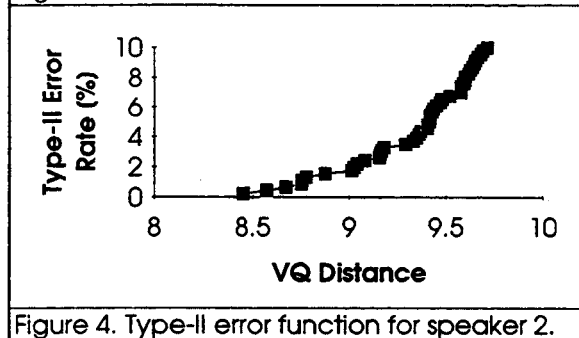


Figure 4. Type-II error function for speaker 2.

### 4. CONCLUSIONS

This study has shown that interspeaker distances in a speaker verification system based on a variance-weighted average VQ distance measure depend critically on F0 differences between impostor and customer. It is further shown that this dependence is asymmetric with large distances between high-pitch impostors and low-pitch customers while no such large increase in VQ distance was found between low-pitch impostors and high-pitch customers.

As a practical consequence, it was demonstrated that the important low error range of the type-II error function for a customer can be determined from a moderate sample of impostors having fundamental frequencies similar to that of the customer.

### ACKNOWLEDGEMENT

This research has been carried out on behalf of the Harry Triguboff AM Research Syndicate.

### REFERENCES

- [1] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, Vol.9, pp351-356, 1990.
- [2] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, Vol.COM-28, pp84-95, 1980.