A LOGNORMAL TIED MIXTURE MODEL OF PITCH FOR PROSODY-BASED SPEAKER RECOGNITION

M. Kemal Sönmez Larry Heck Mitchel Weintraub Elizabeth Shriberg

SRI International
333 Ravenswood Ave. Menlo Park, CA 94025

ABSTRACT

Statistics of pitch have recently been used in speaker recognition systems with good results. The success of such systems depends on robust and accurate computation of pitch statistics in the presence of pitch tracking errors. In this work, we develop a statistical model of pitch that allows unbiased estimation of pitch statistics from pitch tracks which are subject to doubling and/or halving. We first argue by a simple correlation model and empirically demonstrate by QQ plots that "clean" pitch is distributed with a *lognormal* distribution rather than the often assumed normal distribution. Second, we present a probabilistic model for estimated pitch via a pitch tracker in the presence of doubling/halving, which leads to a mixture of three lognormal distributions with tied means and variances for a total of four free parameters. We use the obtained pitch statistics as features in speaker verification on the March 1996 NIST Speaker Recognition Evaluation data (subset of Switchboard) and report results on the most difficult portion of the database: the "one-session" condition with males only for both the claimant and imposter speakers. Pitch statistics provide 22% reduction in false alarm rate at 1% miss rate and 11% reduction in false alarm rate at 10% miss rate over the cepstrum-only system.

1. INTRODUCTION

Statistics of pitch have recently been used as prosodic features in speaker recognition systems with good results and have proven to be more robust than cepstra to acoustic environmental mismatches [1]. Also, in the context of objective measures of speaker recognizability, pitch statistics have been shown to be among the best descriptors, mean pitch being the most descriptive feature [2]. In this work, we address the problem of unbiased estimation of pitch statistics using data from imperfect pitch trackers subject to halving/doubling. Prior work [1] has assumed a Gaussian distribution for clean pitch and addressed the pitch doubling/halving by hard thresholded outlier elimination. Pitch histograms, however, reveal the skewed nature of the distribution. In this work, we argue via a simple correlation model for pitch and show by quantile plots that the clean pitch has a *lognormal* distribution, *i.e.* the logarithm of the clean pitch has a Gaussian distribution. Regarding the pitch tracking errors, we propose a probabilistic model of pitch halving/doubling that results in a lognormal tied mixture distribution for the estimated pitch with four free parameters. We use an auditory model based pitch tracker [3] which uses a model of cochlear filtering to compute autocorrelation-like functions and dynamic programming for tracking and voiced/unvoiced decisions. We use the obtained pitch statistics in a parallel cepstrum/pitch speaker verification system and report results on the March 1996 NIST dataset.

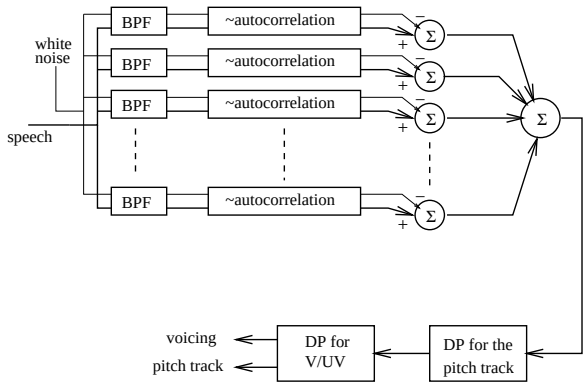

Figure 1. Auditory-model-based pitch tracker

2. PITCH TRACKER

The pitch tracker used in this work (see Fig. 1) is based on a model of cochlear filtering and computes autocorrelation-like functions of the pitch lag for multiple frequency bands followed by two stages of dynamic programing for pitch period estimation and voiced/unvoiced decisions [3].

A summary of key steps in the processing is as follows:
(i) Filtering of the waveform using 14 different bandpass filters to allow pitch tracking of vowels with small residuals or with one very strong formant,
(ii) Computation of autocorrelation-like function using each of bandpass filters in order to preserve/enhance amplitude modulation,
(iii) Subtraction of autocorrelation-like function of white noise from each channel in order to normalize the autocorrelations,
(iv) Summation of different normalized autocorrelation vectors from each channel to obtain the frame normalized autocorrelation function,
(v) Dynamic programming to compute pitch track,
(vi) Computation of maximum "periodicity" at pitch pe-

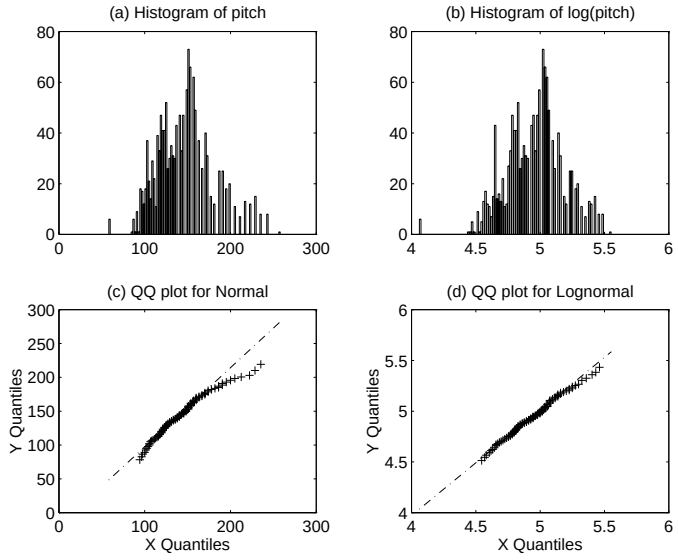

Figure 2. Histograms and QQ plots for Pitch

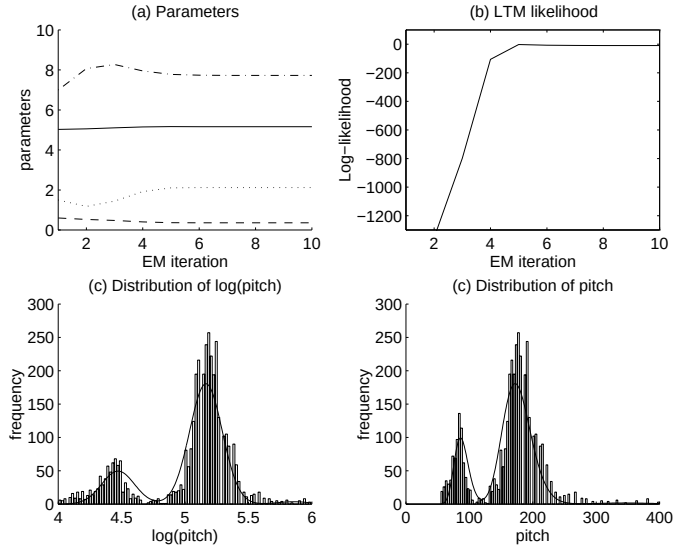

Figure 3. Lognormal Tied Mixture estimation by EM algorithm

riod to allow for rapid amplitude changes,
(vii) Dynamic programming for voiced/unvoiced decision.

3. DISTRIBUTION OF PITCH

Prior work on pitch for speaker identification has assumed that pitch has a normal distribution [1]. Pitch histograms, such as the one in Figure 2(a) for a single female talker, however, indicate that the pitch distribution is skewed. The plot of the Quantiles of the normal distribution vs. the Quantiles of the pitch data (QQ plot) in Figure 2(c) clearly shows (the deviation from the $y = x$ line) that the tails do not match. In the same figure, we show that the lognormal distribution fits pitch histograms much better than the normal distribution, *i.e.*

$$\log(F_0) \sim N(\mu, \sigma^2). \tag{1}$$

Figure 2(b) depicts the histogram for the logarithm of the pitch which is symmetric, and the QQ plot in Figure 2(d) demonstrates a very good fit between the quantiles of the data and those of the lognormal distribution.

We also present the following adaptation of Gibrat's argument [4] for the plausability of the lognormal as the distribution of the pitch. Denote the pitch periods as T_n, $n = 1, 2, \ldots, N$. Given T_n and T_{n-1} are highly correlated, assume a model of the form:

$$T_n = (1 + X_n)T_{n-1}, \;\; n = 1, 2, \ldots, N, \tag{2}$$

where $\{X_n\}$ is a sequence of independent random variables small in magnitude compared to 1, whose distributions are not known. In fact, the dynamic programming formulation for pitch period estimation allows for such a model. Then, T_n can be written as

$$T_n = T_0 \prod_{k=1}^{n} (1 + X_k) \tag{3}$$

which, because X_n is much smaller than 1, simplifies to

$$\log(T_n) \approx \log(T_0) + \sum_{k=1}^{N} X_k \,. \tag{4}$$

One immediately observes by the Central Limit Theorem that $\log(T_n)$ tends to a normal distribution for large n. Therefore, pitch periods are lognormally distributed: $\log(T) \propto N(\mu, \sigma^2)$. Then, pitch, F_0, with the model (2)

$$\log(F_0) = -\log(f_s) - \log(T) \propto N(-\log(f_s) - \mu, \sigma^2) \tag{5}$$

is also lognormally distributed.

4. LOGNORMAL TIED MIXTURE

In the previous section, we demonstrated that lognormal is a suitable distribution for clean pitch, F_0. Now, we consider estimated pitch, denoted by $\tilde{F}_0$, which has been exposed to halving and doubling. We propose the following probabilistic model: Let $\tilde{F}_0 = f(F_0)$ where $\log(F_0) \sim N(\mu, \sigma^2)$ and f is a probabilistic mapping

$$f(x) = \begin{cases} \frac{x}{2} & \text{with prob. } \beta \\ x & \text{with prob. } \alpha \\ 2x & \text{with prob. } 1 - \alpha - \beta \end{cases} \tag{6}$$

This results in the following lognormal tied mixture (LTM) model for the observed pitch distribution:

$$\begin{aligned} \log(\tilde{F}_0) \sim LTM(\alpha, \beta, \mu, \sigma) = \\ \beta \cdot N(\mu - \log(2), \sigma^2) + \alpha \cdot N(\mu, \sigma^2) \\ +(1 - \alpha - \beta) \cdot N(\mu + \log(2), \sigma^2) \end{aligned} \tag{7}$$

Expectation-Maximization (EM) algorithm is used to estimate the parameter vector $(\alpha, \beta, \mu, \sigma)$. Figure 3 shows (a) the parameters and (b) the log-likelihood vs. EM iterations and the obtained models for (c) pitch and (d) log-pitch of a single female talker in Switchboard. Convergence is very fast because of the small number of free parameters in the model.

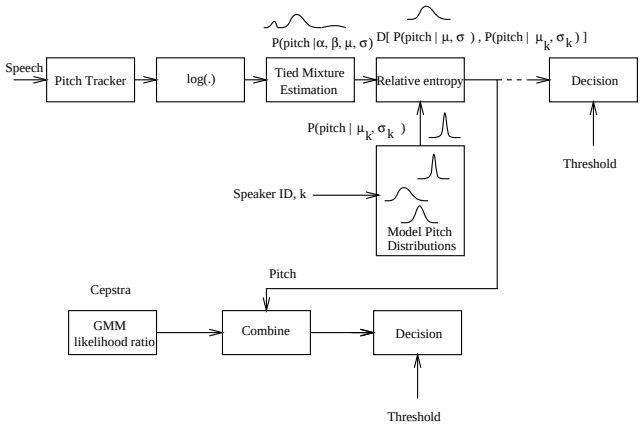

Figure 4. Speaker verification by LTM estimation of pitch statistics

5. SPEAKER RECOGNITION SYSTEM

The block diagram of the system is shown in Fig 4. In the front-end F_0 is computed by the pitch tracker and then transformed to log domain. Tied mixture estimation produces the parameters of the LTM for comparison with the model parameters estimated off-line from a training utterance 2 minutes in duration. The relative entropy between the main (full-voice) modes of the pitch distributions of the model and the utterance is computed as the final score for the pitch system.

The complementary spectral speaker recognition system includes a cepstrum based front-end, EM-trained Gaussian mixture speaker models (GMMs), and a speaker-independent GMM for score normalization. Speech segmentation is accomplished with selection of the top 75% frame-based likelihood ratio scores. Features for the cepstrum-based system are 27 mel-cepstral coefficients computed from a 25 ms window with a frame rate of 10ms, via a filterbank with 28 trapezoidal-shape filters over a warped frequency scale with cepstral mean subtraction over the utterance. The log-likelihood scores of the cepstral system and the relative entropies of the pitch distributions are combined to obtain the overall score.

6. DATABASE

The database we used in our experiments is the March 1996 NIST Speaker Recognition Evaluation. This database is a subset of Switchboard, a conversational-style corpus of long distance telephone calls. The subset consists of 40 claimant speakers (21 male, 19 female) and approximately 400 impostor speakers (200 male, 200 female). There are three training conditions for each claimant speaker: "one-session" (all training data from one phone call, i.e., one handset), "one-handset" (training data from two phone calls, but with one handset), and "two-handset" (training data from two different handsets). Each training condition uses 2 minutes of training speech from the claimant speaker. There are two testing conditions: "matched" and "mismatched" telephone numbers, referring to whether or not the telephone used during testing was the same as that used in training. The results reported in this paper

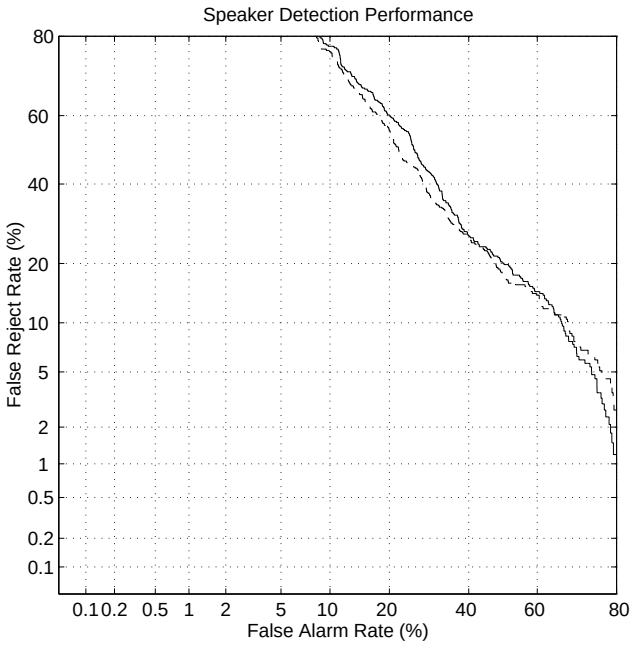

Figure 5. Speaker detection curves for iterative averaging (solid) and LTM estimation (dashed)

Feature Set	1% FAR		10% FAR		EER	
	mat	mis	mat	mis	mat	mis
Pitch	84.8	87.1	47.5	68.1	29.3	32.9
Cepstrum	18.9	73.9	2.7	42.0	5.9	20.2
C + P	14.7	59.9	2.4	34.8	5.3	19.6

Table 1. Speaker verification results

are focused on the most difficult portion of the database: the "one-session" condition with males only for both the claimant and impostor speakers. Both test conditions over the 30-second duration utterances are reported.

7. RESULTS AND FUTURE WORK

The LTM estimation produces a modest but consistent gain over iterative outlier elimination described in [1]. The speaker detection results for both techniques are shown in Fig. 5. LTM uses statistics of halved/doubled datapoints and also produces the ratio of pitch halving to integral pitch. The rate at which a given speaker produces creaky speech (or vocal fry) during spontaneous speech (which is the main reason for pitch halving) is expected to have further discriminating power. Our model produces this rate as the ratio α/β. Future work will include the "vocal fry rate" as an additional feature.

The speaker detection results for the complete system are shown in Figures 6 and 7 for the matched telephone number case (Fig. 7) and the mismatched telephone number case (Fig. 8). Numerical results are reported in Table 1 in terms of false alarm rates (FARs) at 1% and 10% miss rates, and equal error rates (EERs) for both the matched and mismatched cases. There is a remarkable gain in performance in both cases, with the gain in the mismatched case significantly larger: 22% reduction in false alarm rate

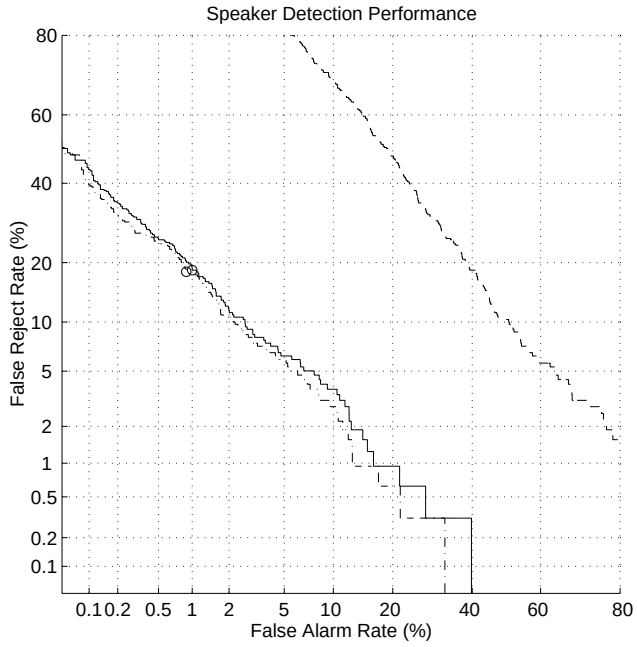

Figure 6. Matched telephone number speaker detection curves: Pitch (top), Cepstrum (middle), C+P (bottom)

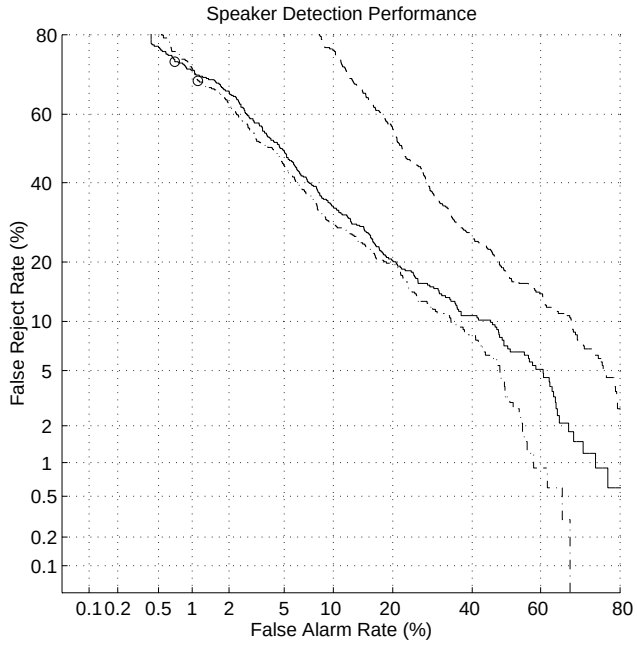

Figure 7. Mismatched telephone number speaker detection curves: Pitch (top), Cepstrum (middle), C+P (bottom)

at 1% miss rate and 11% reduction in false alarm rate at 10% miss rate over the cepstrum-only system. This is a direct result of the fact that pitch is affected much less by the transducer characteristics (carbon-button vs. electret) than cepstrum which is evident from the pitch-only performances in both cases.

REFERENCES

[1] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett. “Robust prosodic features for speaker identification”. In *NIST Speaker Recognition Workshop*, March 1996.

[2] B. F. Necioğlu, M. A. Clements, and T. P. Barnwell III. “Reliability assessment and evaluation of objectively measured descriptors for perceptual speaker characterization”. In *ICASSP*, May 1997.

[3] M. Weintraub. *A Theory and Computational Model of Auditory Monaural Sound Separation*. PhD thesis, Stanford University, 1985.

[4] Johnson N. L., Kotz S., and N. Balakrishnan. *Continuous Univariate Distributions: Volume 1*. Wiley, 1994.