A Study of Automatic Pitch Tracker Doubling/Halving "Errors"

Kathleen Murray
Department of CIS
Moore School Building
200 South 33rd St.
19104 Philadelphia, PA US
kmurray@voicenet.com

Abstract

Manually verified pitch data were compared with output from a commonly used pitch-tracking algorithm. The manual pitch data made statistically significantly better "final rise" predictions than the automatic pitch data, in spite of great similarity between the two sets of measurements. Pitch Tracking doubling/halving errors are described.

Introduction

Automatically captured prosodic information is relevant to both automatic speech recognition and speech synthesis. Pitch information, though regarded as highly relevant, has not been scrutinized in detail as with respect to automatic pitch trackers. This study presents a comparison of hand-verified pitch measurements ("hand") with measurements from a commonly used pitch tracking algorithm ("automatic"), Talkin (1995). For this paper pitch will be defined as the aurally perceived information that loosely correlates with the fundamental frequency of a section of a speech waveform. The organization of this paper is as follows: First, the corpus used is described and justified. The next section describes comparisons of the hand-corrected pitch measurements and the automatic pitch tracker output. Next, results are presented with respect to the detection of utterance-final rises and falls. Lastly, the future work section connects conclusions from this specific study to related work on pitch and perception, and describes discourse-related applications that could benefit from a study of this kind.

1 Corpus Description

The 1992 and the 1993 dialogs from the TRAINS corpus Heeman and Allen (1995) were developed to facilitate the study of collaborative dialogs. In these dialogs, one person guided a "user" through a railroad freight system transportation task, and a monitor recorded the speech without interruption. Trained phoneticians labeled a subset of this speech with ToBI information Beckman and Ayers 1994/1997. Around 26 minutes of speech from a subset of these dialogs were analysed with respect to pitch. "Wedw" software was used Bunnell et al. (1992) by a linguistically trained annotator first in automatic mode. Hand consistency checks then examined glottal pulse locations in a wideband spectrogram. Wedw's wideband spectrogram displays an extremely darkened region where the glottis closes, approximating glottal pulse locations. Hess (1983) recommended use of a wideband spectrogram for manual verification of pitch tracks, but he conceded that wideband spectrograms do not provide sufficient resolution for the eye. In addition to use of a wideband spectrogram, the annotator carefully regarded the shape of the signal waveform, to be sure that glottal pulse locations were labeled consistently with respect to local peaks in the actual speech waveform. These dialogues were chosen for future in-depth investigations of what intonation-based features could be integrated into an automated dialogue system for determining user intentions and generating appropriate system responses.

2 Pitch Tracking Comparison

One concern in automatic pitch tracking is how to handle occasional events where an octave halving appears in the speech signal, but is not readily perceived by a human listener. The algorithm in Talkin (1995) addresses this issue with special constraints on dynamic programming cleanup of the pitch tracker

output. **Figure 1** below illustrates difficulties in making comparisons between pitch trackers in terms of doubling errors. **Figure 1** plots a manually annotated pitch track that ranges from 75 Hz to 189 Hz, and an automatically generated pitch track that ranges from 74 Hz to 102 Hz, for the interval [1.37,1.98] of a 2.5 second utterance. The words of the utterance are "and pick up three boxcars how long is that". A final rise can be heard at the end of the utterance, indicating a user's request for information from the system. The complete ToBI string associated with the utterance is "H* L-L% L* H-H%".

The last voiced section of utt10 shows the speaker vacillating between one octave and another, but the last ToBI string associated with the utterance is "H-H%", meaning a high phrase accent followed by a high boundary tone. It would be surprising for the speaker to be speaking in the 90-100 Hz range reported by the pitch tracker, because the previous section of speech is actually an octave higher, in the 200-235 Hz range. An octave pitch drop would not make sense in the context of a combination of high ToBI labels. The speaker is female. Initial comparisons are difficult because neither method precisely specifies the pitch information, so no pitch gold standard could be produced without significant manual verification of context-dependent doubling rules. When a section of speech appears to be halved in pitch, that halving could be a perceptually significant drop, or it could be a pitch tracker error.

For the 320 utterances used in the evaluation (see **Section 3**), it was determined that roughly 40,419 10 ms frames had occurred where both methods predicted a voiced frame. When the ratio was taken of X/Y, where X was the automatic measurement, and Y, the hand measurement, it was the case that 96% of the time, this ratio was between .8 and 1.2, meaning that the automatic measurement was 20% off the hand measurement for 96% of the relevant cases. The distance of 20% can be used as a goal for past comparisons of pitch tracker outputs with a "gold standard", although some studies have reported an allowance of 30 Hz Niemann et. al (1994). Using the 20% distance, these two methods of pitch look very similar.

For determining halving amounts, one can consider the percentage of time that the ratio of the hand measurement to the automatic measurement was between 1.7 and 2.2. For the roughly 40,000 10 ms voicing-coincident 10 ms frames, .5% of them could be counted as a halving by the automatic pitch tracker for the female speakers, and .4% of the male speech was halved in pitch. One speaker, "JT", female, comprised half of the female pitch measurements, and had a 1% pitch halving rate. One reason these proportions are so small is that the hand-verified data still has some halved data in it, as **Figure 1** shows. For some measurements, pitch halvings are not "errors" at all, because they can directly reflect the information in the speech signal. When speech from the speaker "JT" of **Figure 1** was corrected for halving, 36% of the ratios between the hand-verified and the automatic data were between 1.7 and 2.2.

3 Detection of "Rise"/"Falls"

This section reports the results of applying a simple classification rule with respect to the different pitch methods. The idea comes from Daly (1994). Often, the last label in a ToBI-labeled utterance is a final boundary tone. For 320 utterances, this was the case, and an association was made between the "H%" (high) boundary tone and a "Rise" and the "L%" (low) boundary tone and a "Fall". When the author listened to these utterances, thirteen were ruled out as not contributing a readily perceived tone. This coarse classification is a first approximation towards a perceptually based evaluation of pitch trackers that focuses on a section of an utterance considered linguistically special Pierrehumbert and Hirschberg (1990). The last part of an utterance can signal a user's intention, such as asking a question.

For classifying final tones, firstly the average pitch value for the last voiced region was calculated, "avg_L", and the average pitch value of the remaining voiced regions was calculated, "avg_R". Next, the longest slope for the last voiced section was calculated, "$slope_L$". Where "avg_L" was greater than "avg_R", or "$slope_L$" was positive, a final high tone was classified. Where "avg_L" was less than "avg_R", or "$slope_L$" was negative, a final low tone was classified.

This combination of slope calculations and simple comparisons were an improvement over the method used in Murray (2001). No other study of this magnitude (the hand labelings yielded roughly 100,000 data points) has been published that combines wideband spectrograms and signal shape to hand measure pitch tracks of conversational speech. **Section 2** showed that for many cases, the outputs of the methods are similar. The hand-verified data could be used to closely examine contexts where a pitch tracker predicts a subharmonic of the perceived pitch. More sophisticated tone classification rules besides this preliminary one could be developed once the accuracy of pitch measurement on conversational speech has been improved.

Table 1 below shows results of this simple classification with respect to hand-verified pitch measurements and automatic ones, and pvalues from a paired t-test. Overall, the hand verified measurements performed better in predicting rises and falls at a $p<.001$ level of significance. The preliminary classification rule used slightly favored female speech over male speech.

4 Future and Related Work

A further step would be to coordinate descriptions of pitch tracking errors with respect to categorizations of laryngealization, such as that of Batliner et al. (1993). A pitch value that is in a "subharmonic" or a "diplophonic" laryngealization, (from MÜSLI) may need to be doubled, and context-dependent doubling rules could make use of the MÜSLI classification. Different kinds of final tone classification can be investigated, once the post-processing of pitch measurements has been better established. Murray (2001) used automatic doubling rules, and a different classification scheme, resulting in lower performance than this study.

Shriberg (1999) mentions laryngealizations in the context of "cut-off" words, *ie*, those words that a speaker did not complete. In a corpus of human-computer dialogues on air travel planning (ATIS), cut-off words had a form of laryngealization corresponding to creaky voice usually on the last 20-50 ms of the word. Better recognition of glottal pulses may lead to improved recognition of cut-off words, which are difficult phenomena for an ASR system. Brøndsted (1997) reported that for a specific dialect of Danish, the presence of a glottal consonant "stød" can cause a pitch tracker to incorrectly report a halved value. Further use of wideband spectrograms to facilitate conventions of locations of glottal pulses and their influence on perceived pitch could assist dialogue research for other languages that have glottalized consonants. Black and Campbell (1995) presented a model for generating intonation patterns based on high-level discourse features automatically extracted from dialogue speech. One particular discourse act label, the so-called "d-yu-Q" label, was reported to rise up to significantly higher pitch values than other discourse act labels. Once pitch halvings and doublings are better understood, additional relationships between pitch changes and discourse acts might be discovered. Lastly, it would be useful to compare this data to outputs of other pitch trackers, such as that of Praat, Paul Boersma and David Weenink. (2001), or an updated version of "EDWave" Bunnell (2001). More sophisticated mathematical models would be interesting to use for the final tone classification, especially with respect to different kinds of pitch tracking algorithms.

5 Conclusions

A task-oriented conversational speech database was manually annotated for pitch, but work remains to make the database precise enough for intonation research. This work focussed on potential halving and doubling errors of pitch trackers, and on evaluation of pitch trackers with respect to a final boundary tone classification. Statistically significantly better classification results were achieved with manual verification of pitch data based on wideband spectrograms and speech waveform information. These results were achieved even though the hand measurements appeared to be very similar to the automatic measurements. Based on the very preliminary results of this study, the following two conclusions can be made at this time: one, that automatic pitch measurements still might not be as accurate as needed in order to make generalizations about intonation contours in conversational speech; and secondly, that the combination of a wideband spectrogram and signal shape is a useful starting point for

creating large-scale hand-verified pitch tracks of conversational speech.

Acknowledgements

My thanks go to James Allen and Lucian Galescu for their support of the corpus annotations. This material is based upon work partially supported by the National Science Foundation grant number IRI-9711009. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

A. Batliner et. al. (1993) MÜSLI: A Classification Scheme For Laryngealizations. In D. House and P. Touati, editors, Working Papers, Prosody Workshop, pp. 176-179, Sweden.

Beckman, M.E. & Ayers Elam, G. (1994/1997) "Guide to ToBI Labelling – Version 3.0", electronic text and accompanying audio example files available at http://ling.ohiostate.edu/Phonetics/E_ToBI/etobi_homepage.html.

Black, A. and Campbell, N. (1995) "Predicting the intonation of discourse segments from examples in dialogue speech", ESCA workshop on spoken dialogue systems, Denmark.

Paul Boersma and David Weenink. (2001) Praat Tool, Institute of Phonetics Sciences of the University of Amsterdam.

Brøndsted T. (1997) "Intonation Contours "distorted" by Tone Patterns of Stress Groups and Word Accent", Intonation: Theory, Models and Applications, Athens (Athanasopoulos).

Bunnell H. T., and Mohammed O. (1992) "EDWave - A PC-based Program for Interactive Graphical Display, Measurement and Editing of Speech and Other Signals." Software presented at the 66th Annual Meeting of the Linguistic Society of America.

Bunnell (2001) "Wedw" pitch tracking software, http://www.asel.udel.edu/speech/Spch_proc/software.html.

Daly N. (1994) "Acoustic-Phonetic and Linguistic Analyses of Spontaneous Speech: Implications for Speech Understanding", PhD thesis, Department of Electrical Engineering, Massachusetts Institute of Technology.

A. Hagen, S. Shattuck-Hufnagel and E. Noeth, (1999) "A Study on Glottalizations and their Automatic Detection". ICPhS Workshop on Non-Modal Vocal-Fold Vibration and Voice Quality Poster Session, San Francisco.

Heeman P.A. and J.F. Allen (1995) The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.

Hess W. (1983) Pitch determination of speech signals : algorithms and devices. New York: Springer-Verlag.

Murray K. (2001) A Corpus-Based Approach Towards Automatic Correction of Pitch Tracker Errors, Proceedings of the NAACL Workshop on Adaptation in Dialogue Systems, Pittsburgh, PA.

Nöth E. et. al. (2000) Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System. TransSAP, 8(5):519-532.

H. Niemann et. al. (1994). Pitch Determination Considering Laryngealization Effects in Spoken Dialogs, Proceedings of ICNN, Vol. 7: 4457-4461 Orlando.

Pierrehumbert, Janet, and Julia Hirschberg. (1990) The meaning of intonational contours in discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (eds.), Intentions in Communication. Cambridge, MA: MIT Press.

Shriberg E. (1999). Phonetic Consequences of Speech Disfluency. Symposium on The Phonetics of Spontaneous Speech (S. Greenberg and P. Keating, organizers), Proc. International Congress of Phonetic Sciences, Vol. 1: 619-622, San Francisco.

Talkin D. (1995) "A Robust Algorithm for Pitch Tracking (RAPT)", from Speech Coding and Synthesis, Kleijn, W.B., Paliwal, K.K. ed. Amsterdam, the Netherlands: Elsevier, 495-518.

Table 1: Final Rise/Falls: %Correct, pvalues

Type (Total)	Hand	Automatic	pvalue
Male (258)	76	68	0.01
Female (49)	82	69	0.06
Overall (307)	77	68	0.001

Figure 1: Pitch Plot of Utt10/d93-20.1, X axis is time in seconds, Y axis is frequency in Hz, squares are automatic measurements, diamonds, hand measurements, time ranges from 1.34 – 1.98 s