A pitch extraction reference database.

Conference Paper · January 1995

Source: DBLP

3 authors, including:

Georg F Meyer
University of Liverpool
95 PUBLICATIONS **1,037** CITATIONS

SEE PROFILE

A PITCH EXTRACTION REFERENCE DATABASE

F. Plante(+), GF Meyer(*), W.A. Ainsworth(+,*)

fabrice@cs.keele.ac.uk, georg@cs.keele.ac.uk, w.a.ainsworth@keele.ac.uk
(+) Dept of Communication and Neuroscience, (*) Dept of Computer Science
Keele University, Keele, Staffs ST5 5BG, UK

ABSTRACT

Many pitch extraction algorithms have been proposed in the past. The comparison of these algorithms is difficult because each study tends to be carried out on a unique data set. The purpose of this project is to develop a database for the comparison of these algorithms. This database is based on a core speech module and several additional modules. The core module contains speech and laryngograph data for 15 speakers reading a phonetically balanced text. A voiced/unvoiced reference file is provided with the speech data. Currently a psychophysics module is available to test the performance of pitch extraction stages on commonly used pitch perception stimuli. The database is intended to be open: contributions and remarks can be send to *georg@cs.keele.ac.uk.*

1. INTRODUCTION

Pitch extraction as a problem is as old as speech processing. Over the past 30 years a number of fundamentally different approaches have been proposed to extract the pitch of speech. All areas of speech research (analysis, recognition, coding, synthesis, perception, pathology) need more or less information about pitch such as voicing decision, fundamental frequency estimation, glottal closure instant detection or glottal wave reconstruction.

The complexity of the task is expressed in the number of methods and algorithms currently available. It is obvious that "none of them work perfectly for every voice, application and environmental condition" (Hess, 1983 [1]). Nevertheless it is important to evaluate the different algorithms for specific applications. Currently, such an evaluation is difficult because evaluations are typically restricted to limited sets of algorithms on proprietary data sets [2, 3, 4, 5].

One of the reasons for this is that research on pitch extraction tends to be 'algorithm-driven' rather than data driven. Details of the algorithms are made public so that they can be re-implemented but little data is shared.

Recent developments in speech recognition research shows how much benefit a common data set can have (DARPA evaluations). The evaluations also show that an approach based on common data has its problems, for instance algorithms can be optimised to perform well on a very restrictive task.

2. AIMS

We propose to make available a common database of speech and psychophysical stimuli. This would allow a systematic evaluation of the strengths and weaknesses of the wide range of pitch extraction algorithms.

Pitch extraction is applied to problems ranging from perceptual models to synthesis applications. It is clear that the requirements are problem specific [1]. A number of core requirements were identified the database should be open, easily obtainable and practical.

Openness

The most important requirement of the database is that it is open. The aim is to provide a small general purpose core database to which more modules, for instance databases used in previous evaluations, or more problem specific data sets can be added (pathological speech, telephone speech, noise, etc). The range of voices should be large enough to allow useful performance measures.

Availability

The database is stored on an ftp server *ftp.cs.keele.dc.uk* and available by anonymous ftp from *pub/pitch.*

Users are able to choose which of the modules will be used for evaluation.

Originally published at: ESCA EUROSPEECH '95 4th European Conf on Speech Communication and Technology, Madrid, September 1995, ISSN 1018-4074, pages 837-840

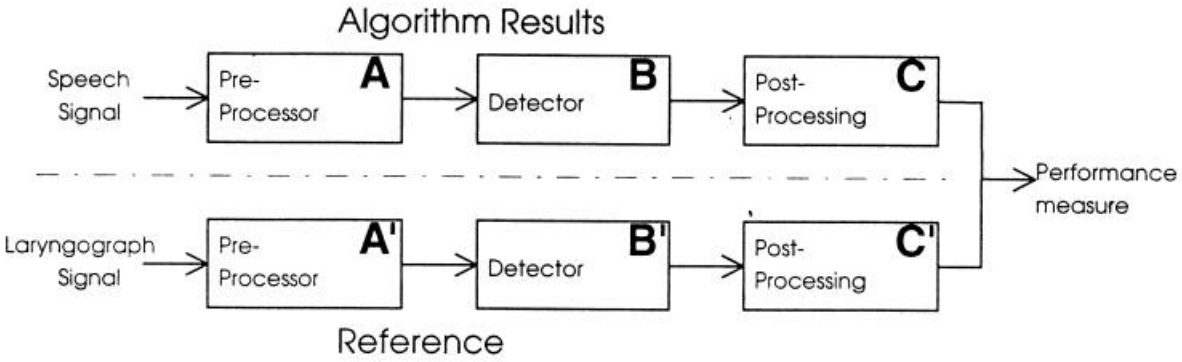

Figure 1: *Schematic diagram of a pitch extraction algorithm. Speech signals are pre-processed to enhance FO related information (A), this information is fed into a detector module (B) to obtain F0 values. Many systems also contain a post-processing stage (C) to smooth out eventual wrong estimates. Reference figures, obtained from a laryngograph trace, also depend on the preprocessor used (A') Ideally the reference and the pitch estimate should be obtained with the same technique.*

Practicality

The database is intended to be used for evaluation rather than optimisation. This means that the size of the core data set and the different modules can (and should) be kept relatively small. New modules will be developed according to user requirements.

3. SPEECH CORE MODULE

3.1. Signals

The core data consists of a phonetically balanced text, 'The North Wind Story", read by 15 native English speakers : 5 mature males, 5 mature females and *5* children (3 males, 2 females). The age and duration of the text are summarised in table 1.

Group	Age	duration (s)
Males	21-60	27-40
Females	20-37	30-38
Children	8-12	30-50

Table 1: Age and duration range for each group of speakers.

The adult readers were recorded in a soundproof room, the children were recorded in a quiet office environment to minimise stress.

The speech data was recorded simultaneously with a laryngograph [6] using a head mounted microphone and a DAT recorder. Both traces were digitised at 20KHz with 16 bit resolution.
The raw data corresponding to the signals were saved in the SAM format: new file extensions are *.pes for the speech signal and *.pel for the laryngograph signal.

A coarsely labelled file (*.pet) containing the beginning and end of each phrase with its orthographic transcription is provided.

3.2. Reference data

Reference files based on the laryngograph signal are provided as the primary reference. It is clear that a laryngograph reference only has application in a subset of algorithms: those concerned with fundamental frequency (FO) estimation and glottal closure extraction.
There are two main problems with the reference:

1) The term 'pitch' implies a perceptual quality, which is known to diverge from the measured F0 for certain stimuli.
2) Whichever algorithm is used to extract an objective pitch estimate from the laryngograph signal will compromise the relative performance of possible pitch extraction algorithms, fig 2. This is due to, for instance, different window sized used in the algorithms and the very different fundamental assumptions made about the signal processed.

For speech data the laryngograph trace, nevertheless is the only feasible hard reference. Users are encouraged to use the laryngograph data to build their own reference data sets.

For convenience a set of more abstract reference files are supplied. These reference files (*.pev) contain a voiced/unvoiced decision and pitch estimates for each 10 ms block of speech in the database. This reference is computed from the laryngograph trace using a floating autocorrelation of 25.6ms duration.

Originally published at: ESCA EUROSPEECH '95 4th European Conf on Speech Communication and Technology, Madrid, September 1995, ISSN 1018-4074, pages 837-840

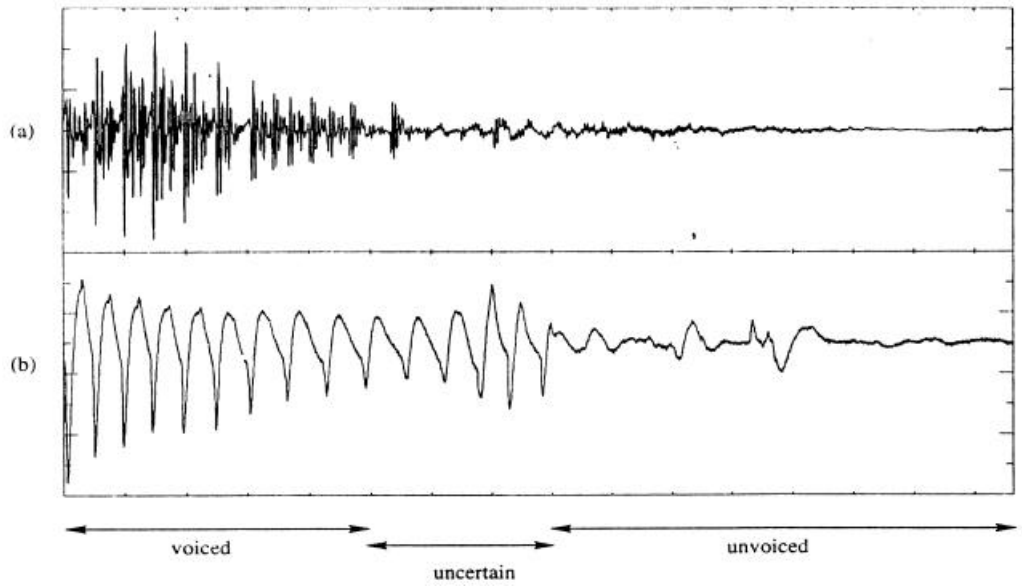

Figure 2: Example of asynchrony between the speech signal (a) and the laryngograph signal (b).

To allow a comparison of algorithms based on different models (production/perception) it is necessary initially to obtain a consistent voicing decision for both speech and laryngograph signals. In theory this should be trivial, in practice there are serious problems. Real recordings contain both sections where the speech waveform looks and sounds voiced while the laryngograph picked up no activity. At the same time laryngeal activity does not always cause a speech waveform. It was decided that segments where no consistent and obvious decision could be made by visual inspection (periodicity in one signal and not in another) are labelled *'uncertain'* using a negative flag. Figure 1 shows an example of these uncertain frames. In this case the periodicity recorded by the laryngograph does not reflect a periodicity on the signal.

For each frame, the reference value is one of the following:

value	a pitch estimate for voiced sections
0	for unvoiced sections
-value	uncertain voiced section (lx data)
-1	uncertain voiced section (voice data)

When periodicity is observed in the laryngograph, but no clear periodic activity can be seen in the speech trace, the corresponding frame is labelled with the negative of the pitch estimate obtained from the laryngograph.
When periodicity is observed in the speech data, but not in the laryngograph the frame is set to *-1* to indicate that it may well be perceived as voiced but that no independent reference exists.
Most of the frames occur at voicing onsets and around plosive bursts. Uncertain frames represent between 2.5 and *5%* of the number of frames, depending of the subject.

The philosophy underlying this decision is that, whichever decisions are taken for a uncertain frame will influence the performance figures for one subset of extraction algorithms, but not others.
If only one source (speech or laryngograph) exists objective evaluation is not possible. The data, consequently, should not be included in comparisons between algorithms.
The data is nevertheless supplied as far as possible to allow maximum flexibility in the algorithm evaluation.
While experimenters using the database are encouraged to use the laryngograph trace to calculate a reference matching the processing performed on the speech data, the labelled uncertain frames should be ignored as far as possible in performance comparisons.
A comparison based on this data is likely to produce good' results because many of the critical frames are removed. It is nevertheless far from clear what, for instance, the perceived pitch of the questionable frames would be.

Originally published at: ESCA EUROSPEECH '95 4th European Conf on Speech Communication and Technology, Madrid, September 1995, ISSN 1018-4074, pages 837-840

4. ADDITIONAL MODULES

4.1. Psychophysics modules

Pitch is defined as a perceptual quality. It could therefore be argued that a production based reference, such as laryngograph traces, is not appropriate and indeed is likely to bias any performance comparisons towards production based algorithms, such as inverse filtering.
Currently no pitch perception data is available for continuous speech. A number of psychophysical stimuli, however, have been proposed and extensively used in the past to explore human pitch perception mechanisms. These are typically very short signals with one or more perceived pitches. The psychophysics module contains the following signals

1) Pure tones
2) Missing fundamental stimulus,
3) Pitch shift sequence,
4) Musical chords,
5) Ambiguous pitch,
6) Amplitude modulated noise
7) Comb filtered noise

The stimuli and models of pitch perception based on them are explained in detail in Meddis and Hewitt [8] or Meyer and Dewar [9].

4.2. Digital Signal Processing Toolbox

A number of DSP tools are available on the same ftp site under pub/DSP. This toolbox contains classical DSP routines (fft, filtering, data generation, etc) and also some pitch extractor algorithms developed at Keele [7]. The reference was computed using this toolbox. The package runs under UNIX and can be used directly as a set of command line commands or can be used with a very simple Xli graphical user interface. The system has been implemented on SUN workstations and tested on a range of other workstations (Silicon Graphics, HP, DEC, LINUX).

5. CONCLUSION

The database proposed here is the first step towards a public database to aid evaluation of pitch extraction algorithms. The database is open and external contributions and remarks are welcome. To keep the size of the database to practical limits, contributions should either cover areas of interest not currently included or have been used previously in algorithm evaluations.
Regular updates will be made. Researchers interested in this database are encouraged to contact the authors by email: *georg@cs.keele.ac.uk.*

6. ACKNOWLEDGEMENTS

The work was supported in part by contract SCI-CT92-0786 of the EC Science Programme.

7. REFERENCES

[1] W. Hess “Pitch determination of speech signals”, Springer-Verlag, Berlin, 1983.
[2] P.C. Bagshaw, S.M. Hiller, M.A. Jack, “Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching.” Proc Eurospeech93, pp. 1003-1006, Berlin 1993.
[3] D.J. Hermes “Pitch analysis” in Visual representations of Speech, Cooke, Beet and Crawford Eds, pp. 1-25, 1993.
[4] L.M. Rabiner, MJ Cheng A.E. Rosenberg, C.A. McGonegal, “A comparative performance study of several pitch detection algprithms”. IEEE ASSP, Vol.24, pp. 399-4 18, 1976.
[5] L. Van Immerseel. J.P. Martens, “Pitch and voiced/unvoiced determination with an auditory model”. JASA,Vo1.9l, pp. 35l1-3526, 1992.
[6] A.J. Fourcin, E. Abberton, “First application of a new laryngograph”. Medical and biological Illustration, Vol.21, pp. 172-182, 1971.
[7] F. Plante, G. Meyer W.A. Ainsworth, “Pitch detection: Auditory model versus inverse filtering” Procs IOA, Vol 16, pp. 81-88, Windermere 1994.
[8] R. Meddis and M.J. Hewitt “Virtual Pitch and Phase sensitivity of a Computer Model of the Auditory Periphery. 1: Pitch Identification” J Acoust Soc AM. 89 pp. 2866-2882, 1991.
[9] G.F. Meyer and I.D. Dewar, “Comparing Pitch Extraction in the Cochlear Nerve and Cochlear Nucleus’. Proc IOA, 16(5), pp. 263-271, 1994.

Originally published at: ESCA EUROSPEECH ’95 4th European Conf on Speech Communication and Technology, Madrid, September 1995, ISSN 1018-4074, pages 837-840