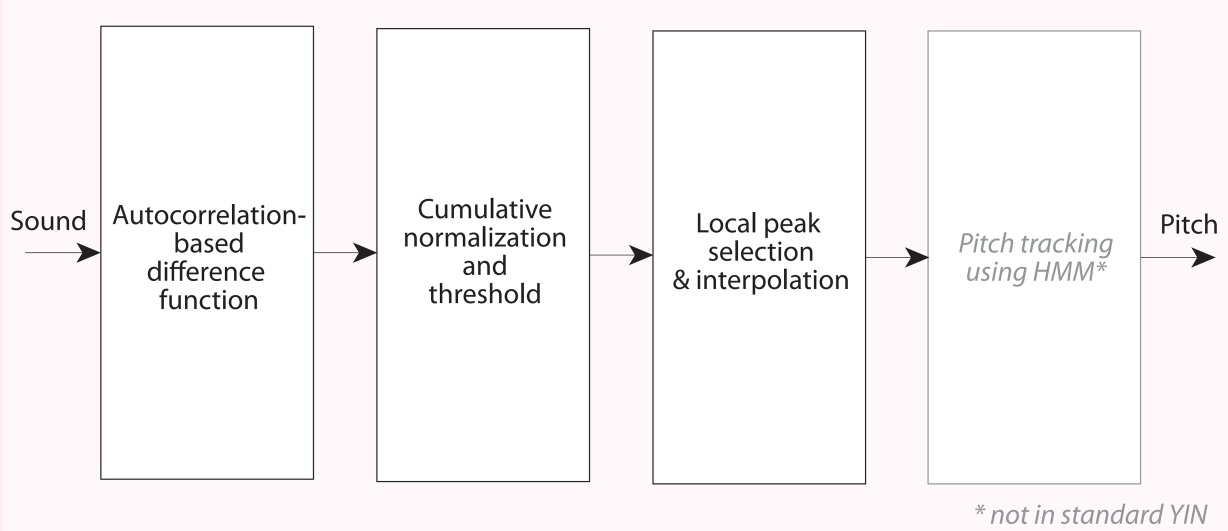
Noise Robust Pitch Tracking by Subband Autocorrelation Classification (SAcC)

Byung Suk Lee & Dan Ellis • Columbia University / ICSI • {bsl,dpwe}@ee.columbia.edu

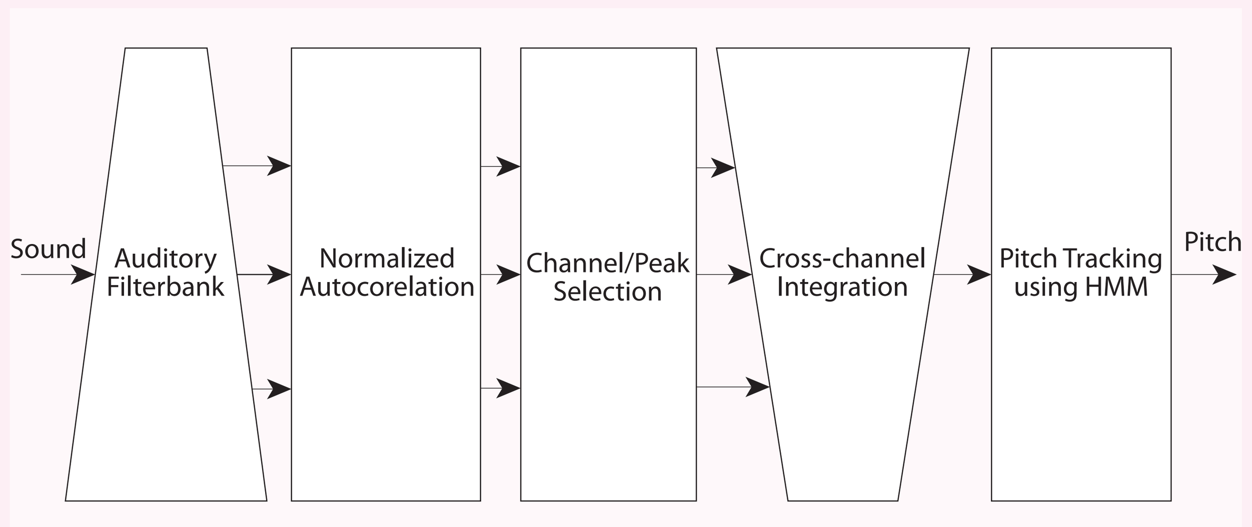
Summary: A neural net classifier is trained to identify the pitch of a frame of subband autocorrelation principal components. Accuracy is greatly improved for noisy, bandlimited speech, matched to the training data.

Background

- "Classic" approaches to pitch tracking reveal time-domain waveform periodicity via autocorrelation or related approaches.
- An example is **YIN** [de Cheveigné & Kawahara 2002]:



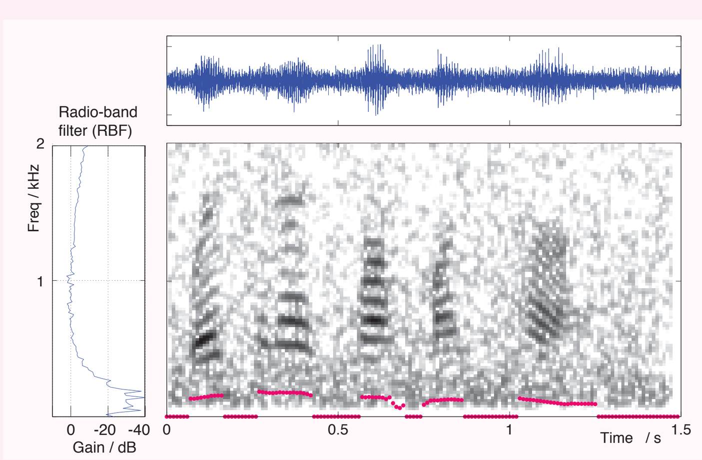
- Added interference hurts the autocorrelation, but filtering into subbands can reveal certain bands with better local SNR. Also, the contribution of each subband can be adjusted in later fusion to select sources.
- An example is [Wu, Wang & Brown 2003] ("the **Wu** algorithm"):



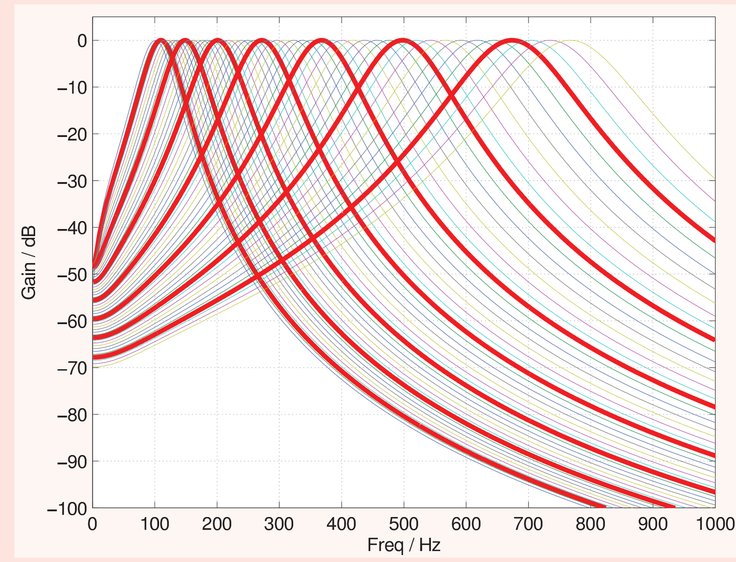
- This work began as an investigation into the peak selection and integration stages of the Wu algorithm. We found that replacing both stages with a trained classifier offered large performance improvements, as well as the chance to train domain-specific pitch trackers.

Material

- We are specifically interested in pitch tracking for low quality radio reception data (e.g. hand-held narrow-FM) (RATS project).
- This data has both high noise/distortion and narrow bandwidth.



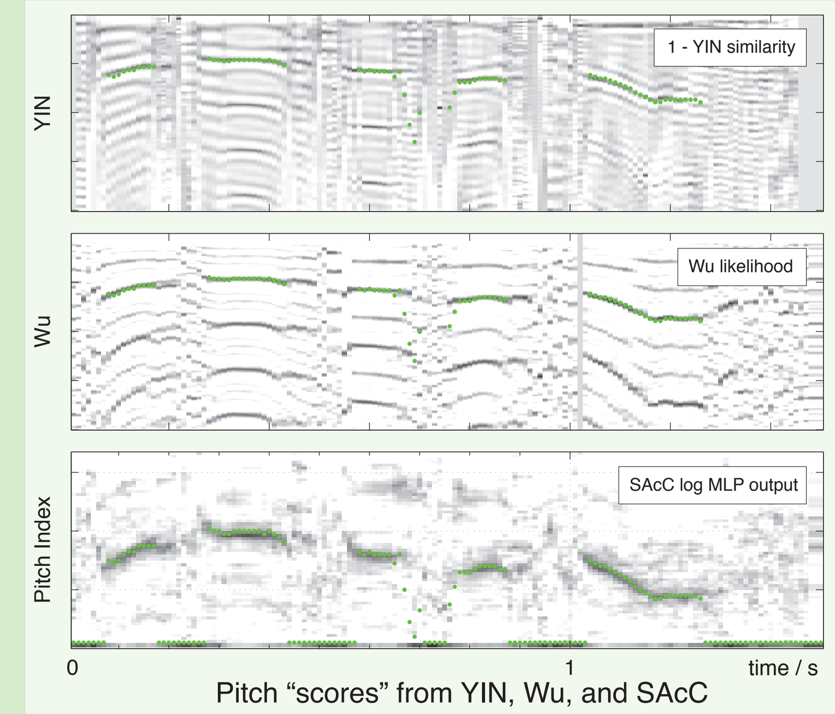
Filterbank



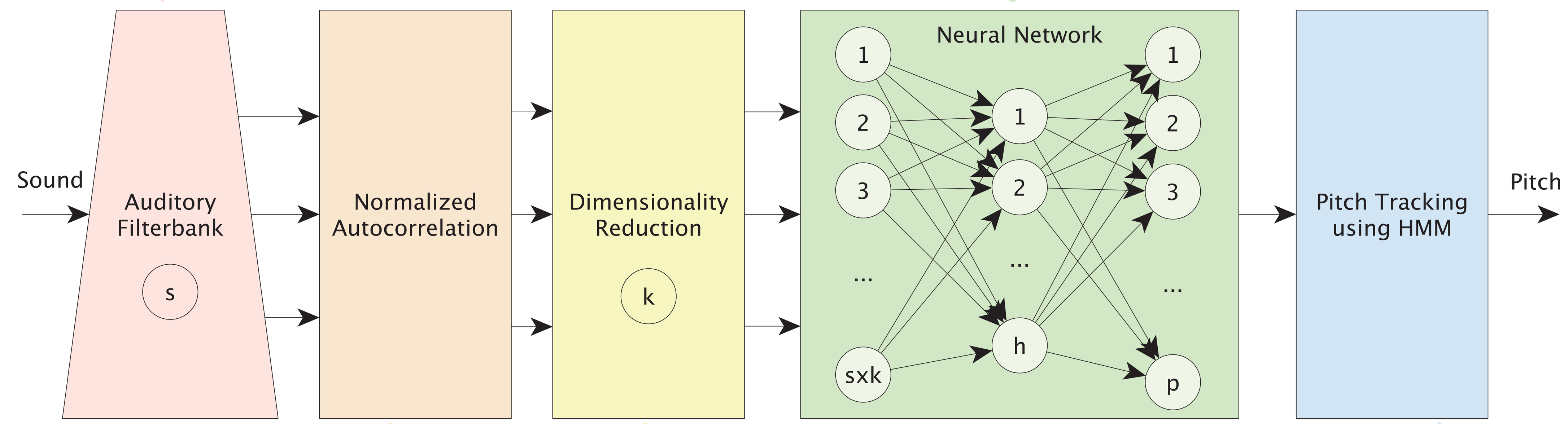
- 48, 4-pole 2-zero ERB-scale cochlea filter approximations form auditory-like subbands.

Classifier

- The core of our system is a trained (MLP) classifier.
- It takes **k** (10) PCA coefficients for each of **s** (48) subband autocorrelations and estimates posteriors over **p** (67) quantized pitch values.
- The MLP is trained discriminatively on noisy data (with ground truth pitches).
- Discriminative training virtually eliminates octave/suboctave errors seen in peak-based pitch tracking.

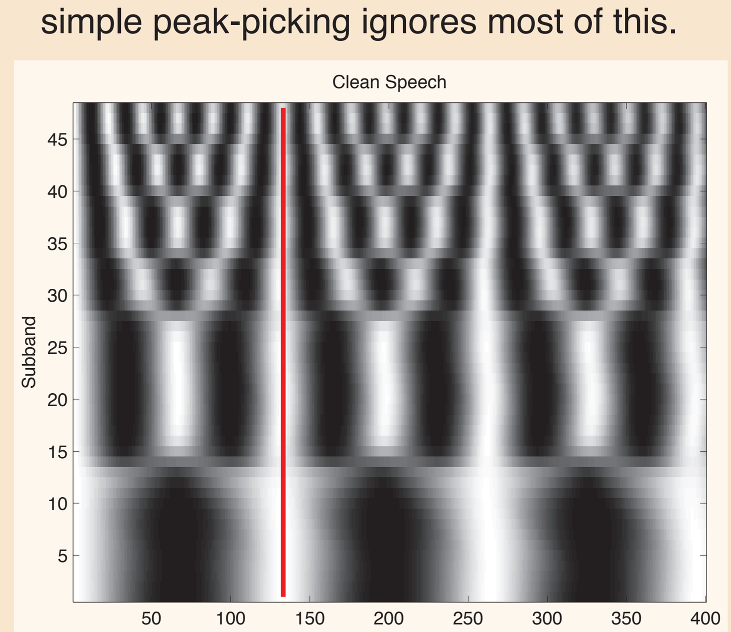


Pitch "scores" from YIN, Wu, and SAcC



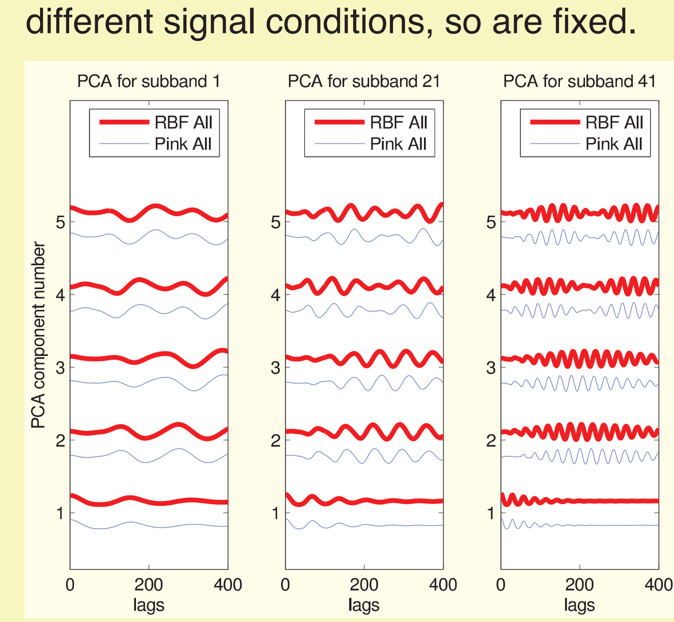
Autocorrelation

- Normalized autocorrelation out to 25 ms (400 samples @ 16 kHz) reveals periodic structure in each subband.
- The autocorrelation in each subband is highly constrained by the bandlimited input.
- Pitch information is distributed throughout each autocorrelation row; simple peak-picking ignores most of this.



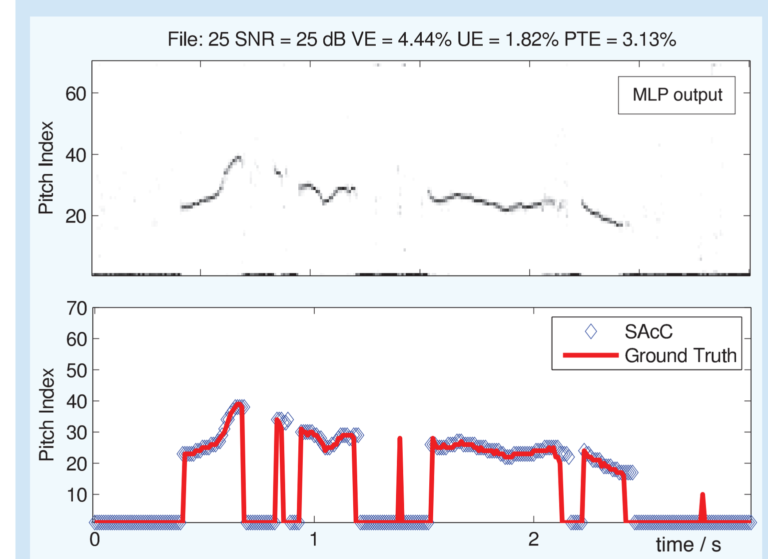
Principal Component Analysis

- The autocorrelations in a given subband always reflect the center frequency, i.e., they occupy a small part of the 400-D space.
- We use per-subband PCA to reduce each band to 10 coefficients, which preserves almost all the variance.
- PCA bases remain stable when learned from different signal conditions, so are fixed.



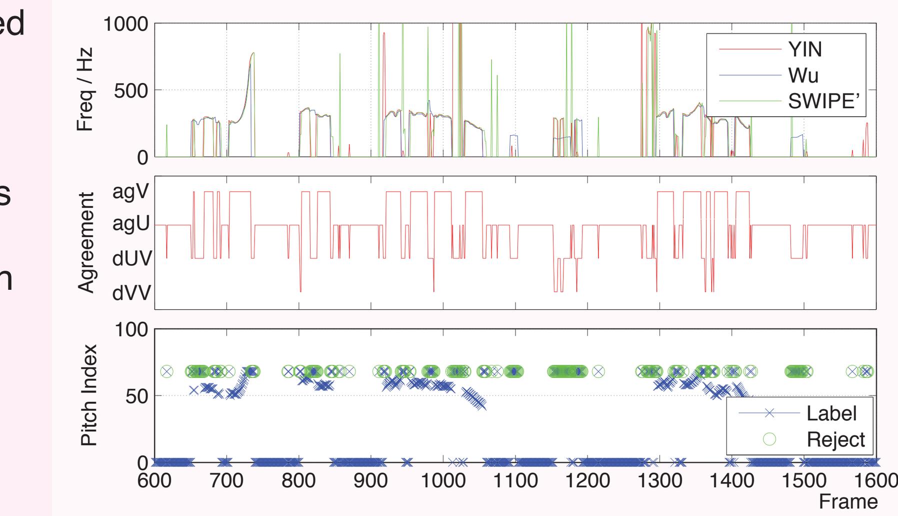
Hidden Markov Model Smoothing

- The MLP generates posterior probabilities across 66 pitch candidates + "no pitch" for every 10 ms time frame.
- These become a single pitch track via Viterbi decoding through an HMM with pitch states.
- Transition probabilities are set parametrically (pitch-invariant) and tuned empirically.



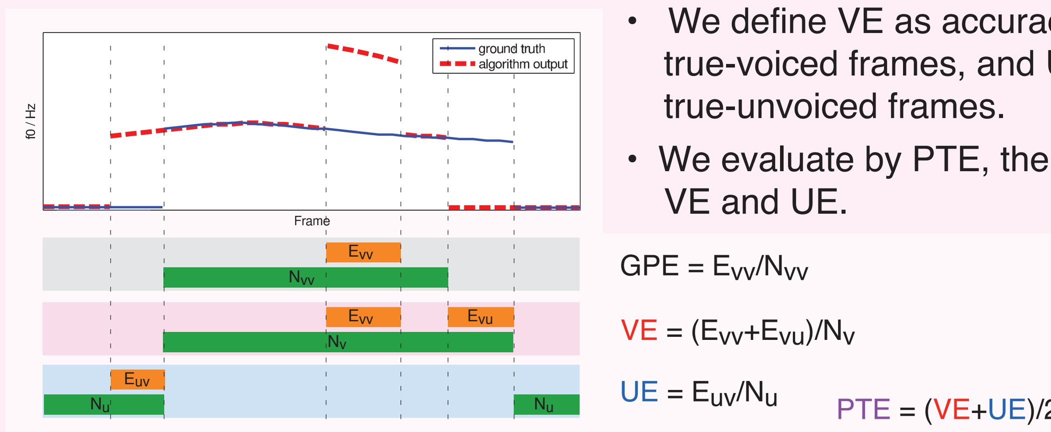
Training Data

- The pitch classifier needs training data with ground truth. Performance improves when training data is more similar to test data.
- We used pitch-annotated data (Keele), then artificially filtered and added noise to resemble the target domain.
- We also generated pseudo-ground-truth for target domain data by using only frames where three independent pitch trackers agreed.



Evaluation

- GPE, the most common pitch tracking metric, only considers frames where both ground truth and system report a pitch value. This rewards "punting" on difficult frames.



- We define VE as accuracy over all true-voiced frames, and UE over all true-unvoiced frames.
- We evaluate by PTE, the mean of VE and UE.

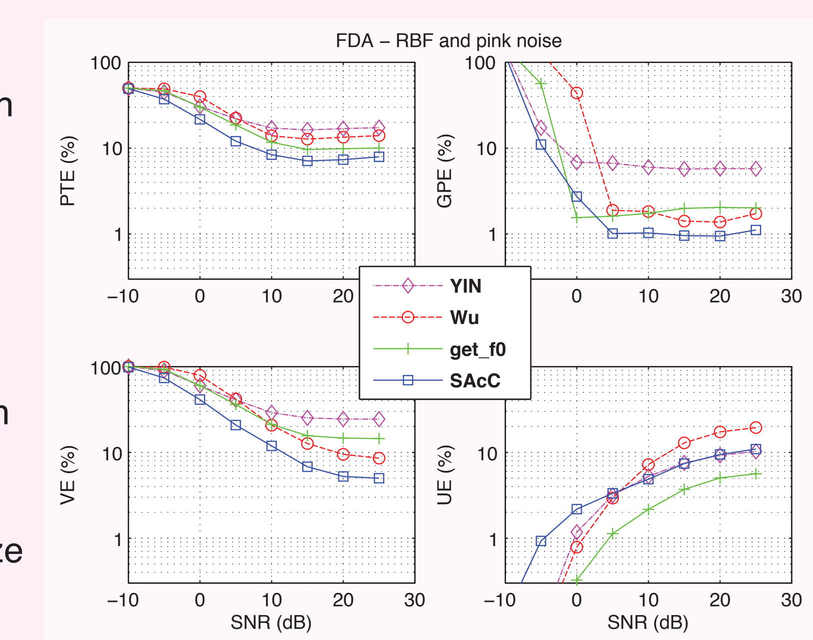
$GPE = E_{vv}/N_{vv}$

$VE = (E_{vv}+E_{vu})/N_v$

$UE = E_{uv}/N_u$

$PTE = (VE+UE)/2$

- We tested on the pitch-annotated FDA data with radio-band filtering and pink noise added at a range of SNRs.
- SAcC trained on Keele data (with similar corruption) substantially outperformed other pitch trackers.
- Later experiments show that SAcC can generalize to mismatched train/test scenarios.



References

A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., 111(4):1917–1930, April 2002.

M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," IEEE Tr. Speech and Audio Proc., 11(3):229–241, May 2003.

MATLAB code available: http://labrosa.ee.columbia.edu/projects/SAcC/