Application of Missing Feature Theory to the Recognition of Musical Instruments in Polyphonic Audio

Jana Eggink and Guy J. Brown
University of Sheffield

Introduction

Instrument recognition can be useful for

- automatic transcription
- automatic indexing
- search for similar music
- query by humming

Computer Instrument Identification

- monophonic -

KD Martin (1999)

- 31 different features, both temporal and spectral
- hierarchical classification scheme
- 27 instruments: 39% isolated tones, 57% phrases
- 6 instruments: 82% phrases

JC Brown et al. (1999, 2001):

- log. scaled cepstral features (MFCCs)
- Gaussian mixture models (GMMs)
- 4 woodwind instruments: average 60%, best 80% phrases

Marques and Moreno (1999)

- log. scaled cepstral features
- support vector machines (SVMs)
- 8 instruments: 70% phrases

Computer Instrument Identification
- polyphonic -

Kashino & Murase (1999)
- time domain approach based on example waveforms
- 3 instruments, specially made recording
- F0s and onsets supplied
- 68% correct, max. polyphony 3

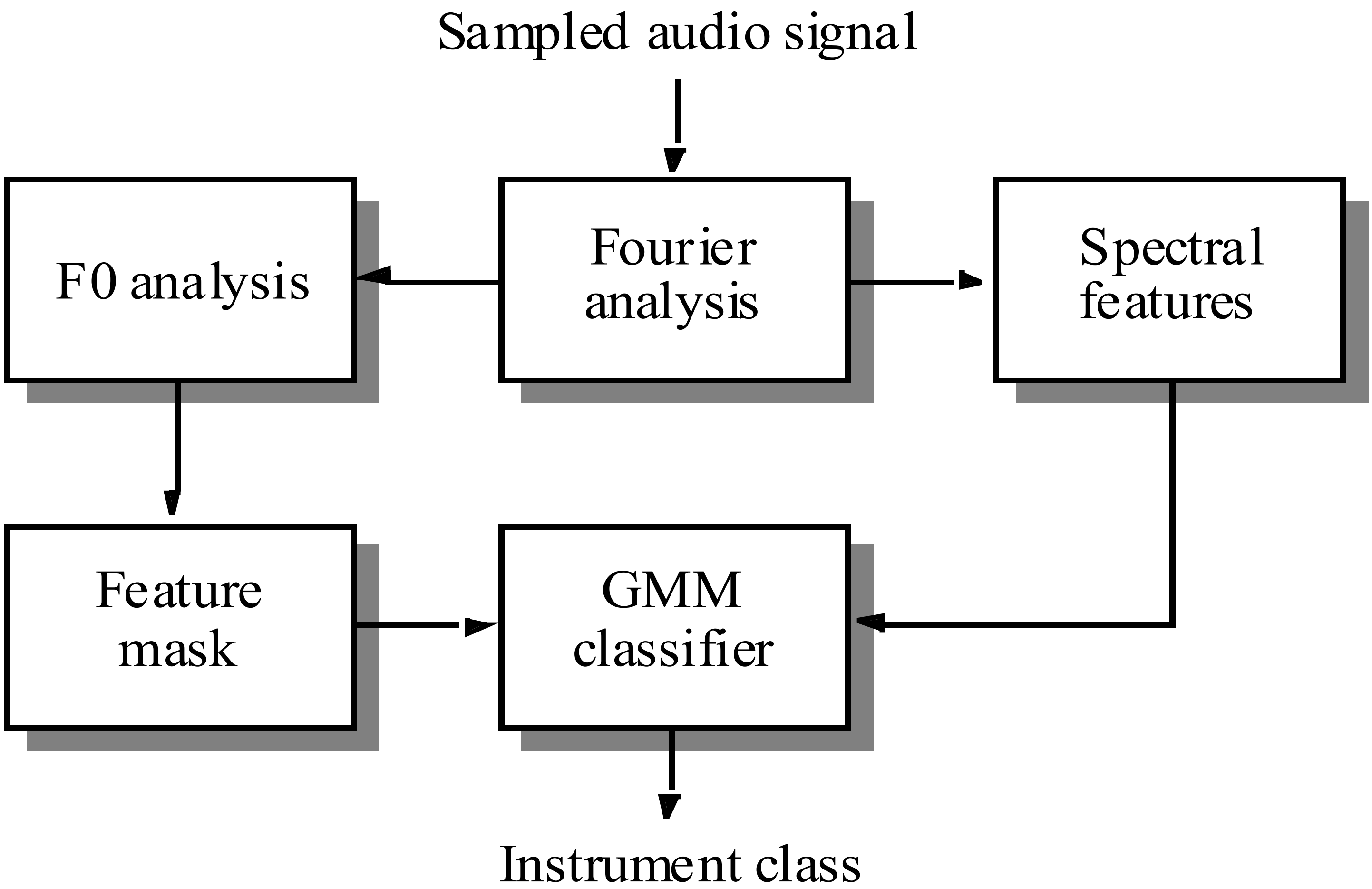
Kinoshita et al. (1999)
- frequency domain approach based on partials, measuring sharpness of onset and spectral energy distribution
- feature values from overlapping partials are (mostly) ignored
- 3 instruments, random 2 tone combinations
- 70% correct (78% if F0s supplied), max. polyphony 2

Our System

Missing feature approach

- sound sources are not only additive, but can also mask each other
- in music, harmonics from one tone often coincide with those of another tone, resulting in energy values that do not correspond to either instrument, therefore
- we exclude features dominated by an interfering sound source from the recognition process,
- resulting in an incomplete, but mainly uncorrupted representation
- classifier is modified to work with partial data

System Overview

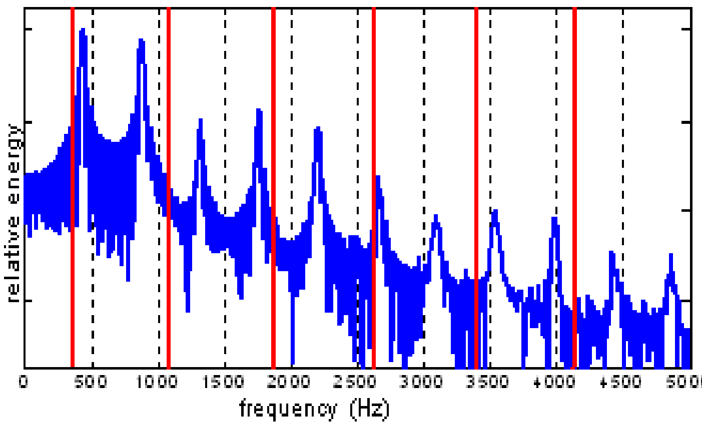


Features

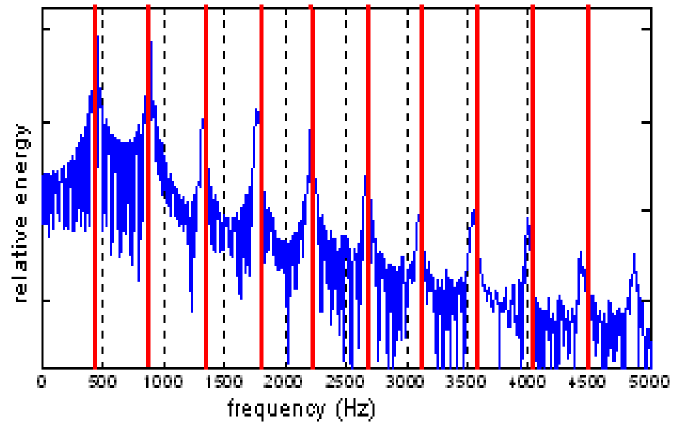
- local spectral features are required for missing feature approach
- frame based (frame length 40 ms)
- energy in narrow frequency bands (60 Hz bandwidth)
- linearly spaced, corresponding to linear spacing of partials
- basically coarse spectrograms

F0-Analysis

• iterative approach based on harmonic sieves (Scheffers, 1983)



bad fitting sieve



best fitting sieve
determines F0

• advantage of direct identification of peaks/harmonics, can be used for more exact mask estimation

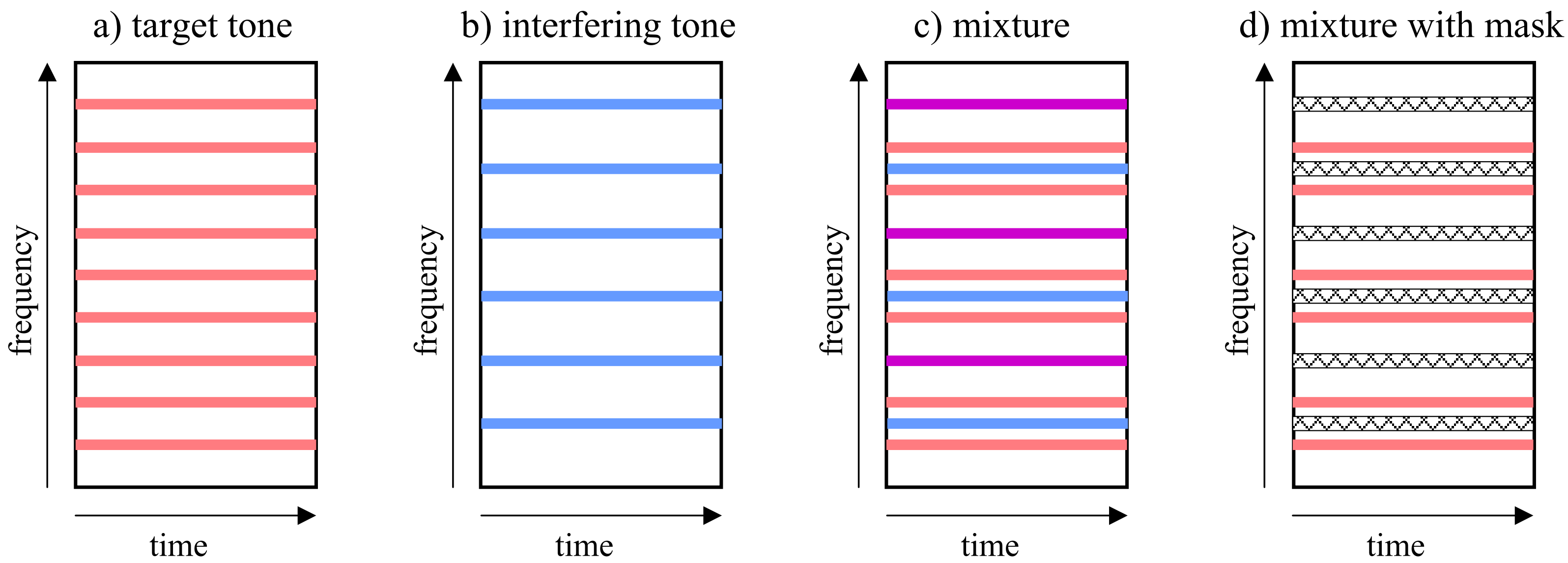
Missing Feature Estimation

- finding reliable and unreliable features is one of the main problems

***a priori* masks**
- do not rely on F0-estimation
- require knowledge of the clean (monophonic) signal
- features are only declared reliable when close (±3dB) to the features derived from the clean signal

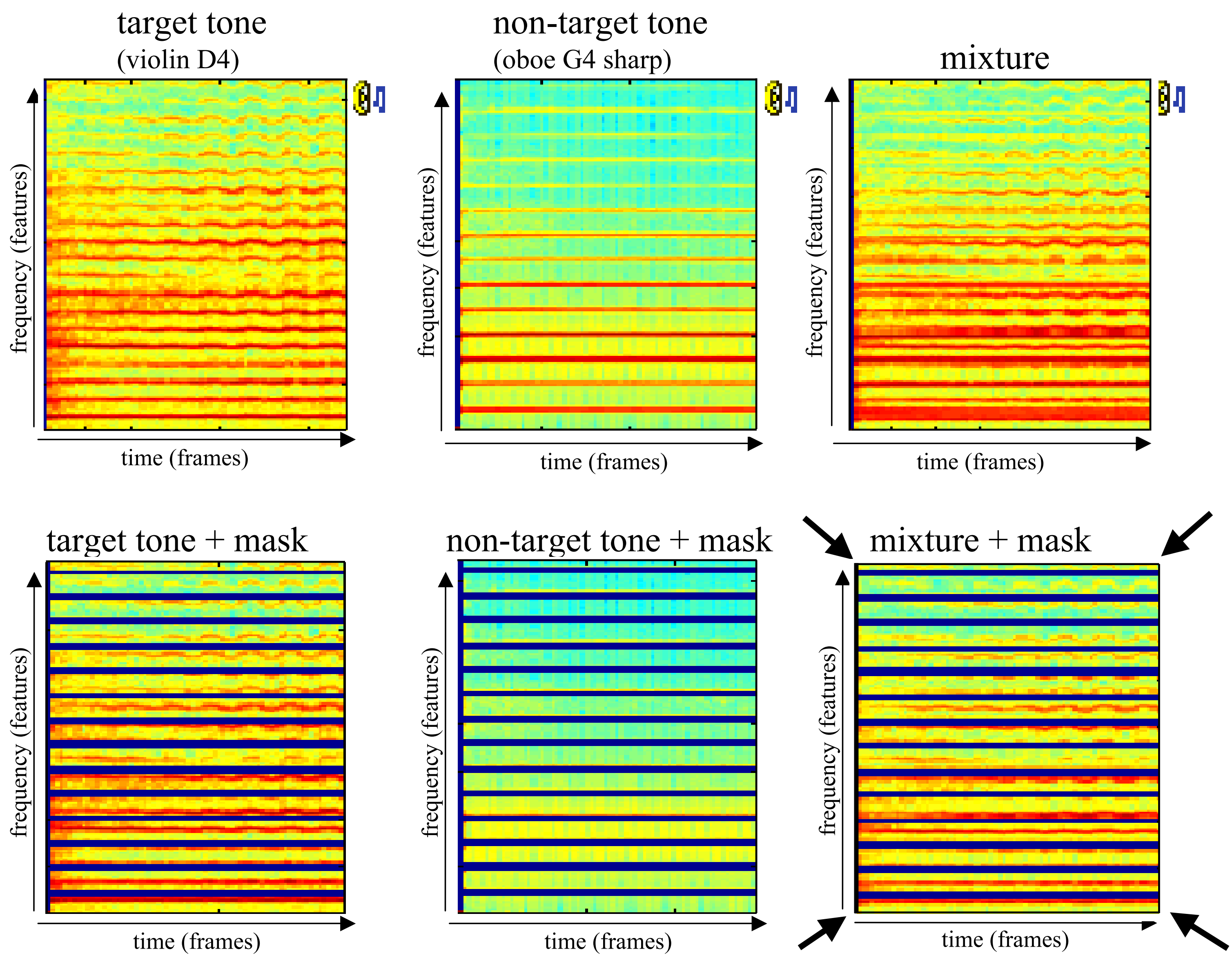
F0-based masks
- instrument tones have an approximately harmonic overtone series
- based on the extracted F0s, all frequency regions where a partial from a non-target tone is found/expected are excluded from the recognition process

Example Features with Mask - I



Simplified spectra of a) the target tone, b) the interfering tone, and c) the mixture of both tones. Energy values which, due to overlapping partials, do not correspond to those of either tone alone are shown in purple. In d) the mixture is overlaid with the mask, represented by hatched bars.

Example Features with Mask - II

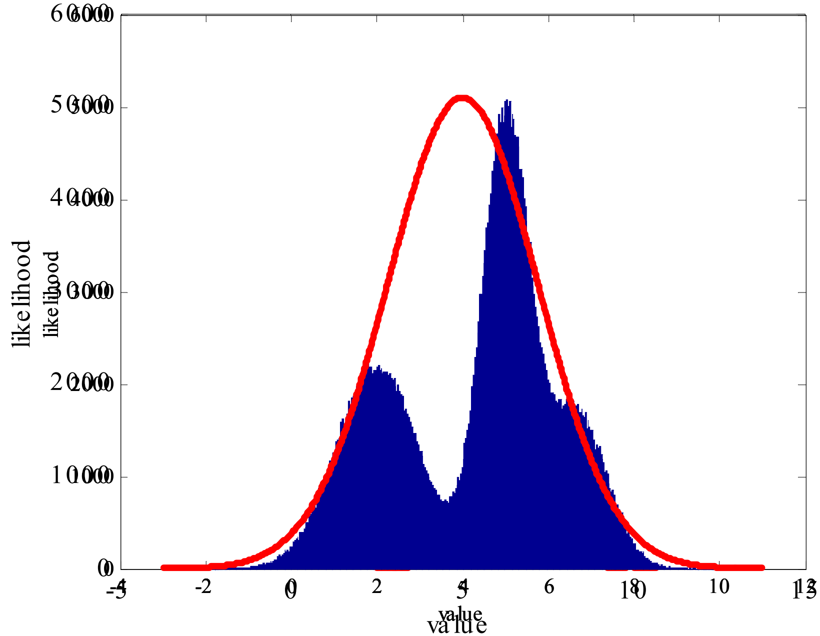


Gaussian Mixture Models - GMMs

- a GMM models the pdf of observed features *x* by a multivariate Gaussian mixture density:

$$pdf(x) = \sum_{i=1}^{N} p_i \Phi_i(x, \mu_i, \Sigma_i)$$

- number of Gaussians has to be chosen manually
- diagonal or full covariance matrices
- means, covariances and mixing coefficients are estimated during training, using
- EM (expectation-maximasation) algorithm



GMMs - Training

- models trained for 5 instruments (flute, clarinet, oboe, violin, cello)
- using both isolated notes and realistic monophonic phrases
- every model has 120 centres and diagonal covariances

GMMs with Missing Features

Missing Features

- unreliable features are ignored, classification is based on reliable features only

Bounded Marginalisation

- unreliable features hold some information, as the observed energy can be regarded as an upper bound for the amount of energy caused by the target signal
- include information from unreliable features by integrating over all possible values below upper bound (i.e. observed energy)

GMMs with Missing Features - math

probability density function (pdf) of observed spectral
D-dimensional feature vector x is modeled as:

$$p(x) = \sum_{i=1}^{N} p_i \Phi_i(x, \mu_i, \Sigma_i)$$

assuming feature independence, this can be rewritten as:

$$p(x) = \sum_{i=1}^{N} p_i \prod_{j=1}^{D} \Phi_i(x_j, m_{ij}, \sigma^2{}_{ij})$$

approximating the pdf from reliable data only leads to:

$$p(x_r) = \sum_{i=1}^{N} p_i \prod_{j \in M'} \Phi_i(x_j, m_{ij}, \sigma^2{}_{ij})$$

N = number of Gaussians in the mixture model, p_i = mixture weight, Φ_i = univariate Gaussians with μ_i = mean vector, m_{ij} = mean, Σ_i = covariance matrix, $\sigma^2{}_{ij}$ = standard deviation, M' = subset of reliable features in Mask M

Bounded Marginalisation - math

$$p(x_r, x_u) = \sum_{i=1}^{N} p_i \Phi_i(x_r, \mu_i, \Sigma_i) \int \Phi_i(x_u, \mu_i, \Sigma_i) dx_u$$

$$\int \Phi_i(x_u, \mu_i, \Sigma_i) dx_u = \frac{1}{2}\left[erf\left(\frac{x_{u,high} - \mu_{u,i}}{\sqrt{2\sigma^2{}_{u,i}}} \right) \right]$$

x_r = reliable features, x_u = unreliable features, $x_{u,high}$ = upper bound of unreliable features

N = number of Gaussians in the mixture model, p_i = mixture weight, Φ_i = univariate Gaussians with μ_i = mean vector, Σ_i = covariance matrix, $\sigma^2{}_{ij}$ = standard deviation

Evaluation

- training and test data always from different recordings

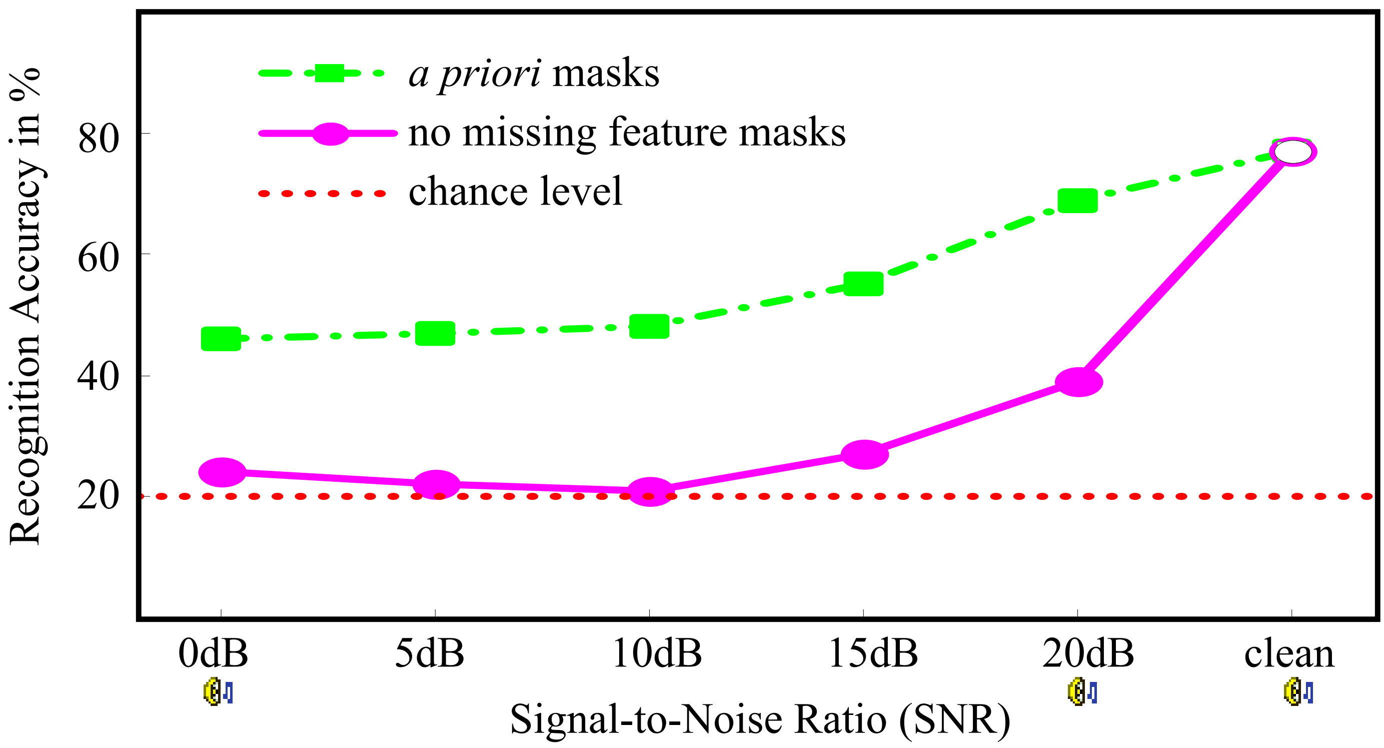
- 3 sample collections (Ircam, Iowa, McGill)
- leave-one-out cross-validation
- only tones from one octave (C4-C5), avoiding cues based solely on the different pitch range of the instruments

- 5 short (2-10 sec) monophonic phrases per instrument, taken from commercially available CDs

Evaluation: Noise

- *a priori* Masks -

- clean, monophonic examples: 77% (samples and phrases)
- mixed with white noise at different SNRs
- missing feature masks improved accuracy by 27% average



Evaluation: 2 simultaneous Instruments

		average	flute	clarinet	oboe	violin	cello
samples	**mono**	**66%**	67%	59%	85%	65%	56%
	a priori	**62%**	73%	47%	73%	68%	51%
	pitch-based	**47%**	54%	44%	48%	64%	31%
phrases	**mono**	**88%**	100%	100%	70%	100%	70%
	a priori	**74%**	72%	49%	63%	89%	94%

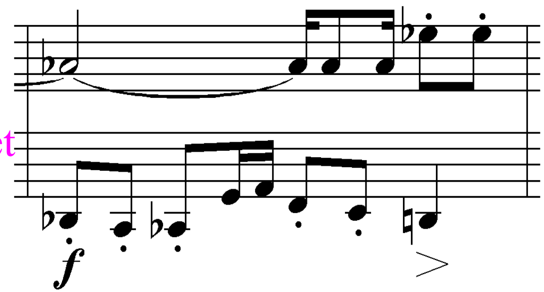
Evaluation: ‘real’ Duet

- duet for flute and clarinet by H. Villa-Lobos
- F0s extracted by the system

original score:

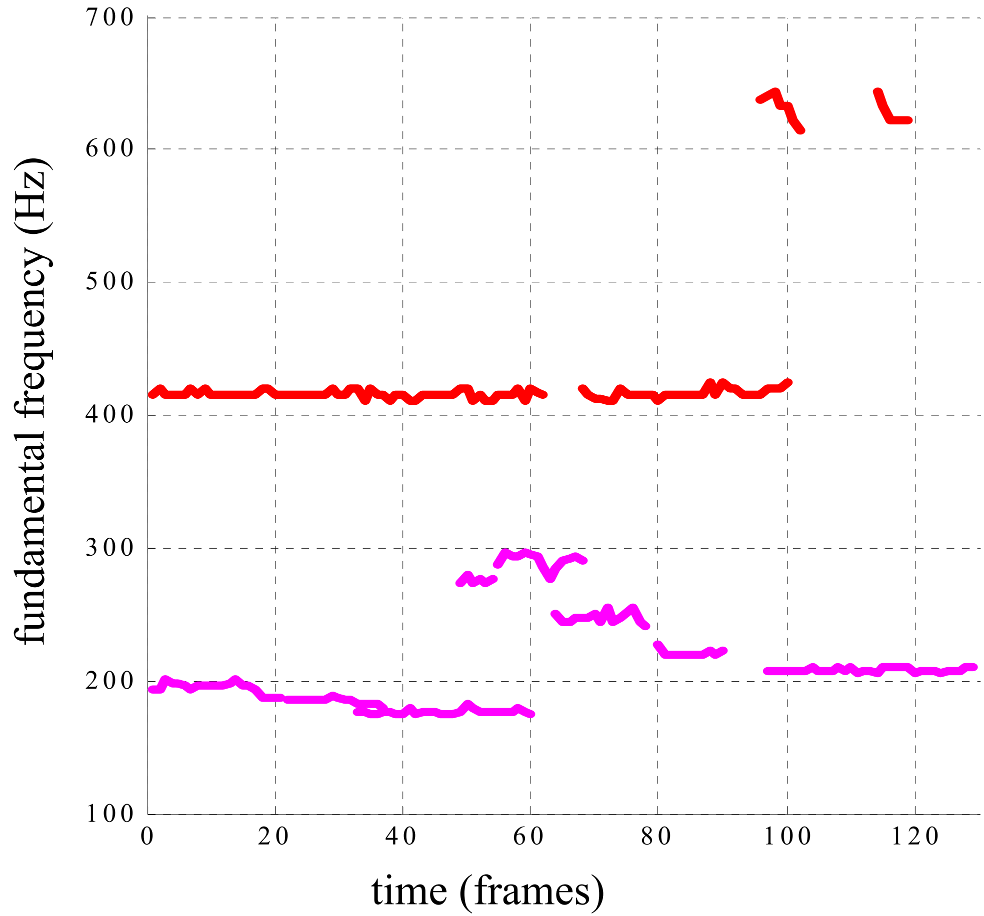
flute

clarinet in A



F0s according to the score in Hz:
415 - 415 - 415 - 622 - 622
208 - 185 - 175 - 277 - 294 - 247 - 220 - 208

system output:



Evaluation: Bounded Marginalisation

- no improvement for combinations of 2 instrument sounds
- strong improvement for instrument sounds mixed with white noise, results as good as the clean condition for all SNR levels
- different energy distribution:
  - a harmonic sound stronly increases energy values in few features
  - (white) noise lightly increases energy values in many features
- bounded marginalisation seems to improve results mainly when the difference between observed and ‘true’ feature value is small
- could be very useful for instrument recognition in noisy, low quality recordings

Conclusions and Future Work

- good results so far, works with realistic stimuli
- some drop between *a priori* and pitch-based masks – more acurate masks needed
- for small ensembles only
- peaks / harmonics are reliable, work on representation that only uses these, train on limited representation

The End

- **Thank you for your attention!**

- Jana Eggink acknowledges the financial support provided through the European Community's Human Potential Programme under contract HPRN-CT-2002-00276, HOARSE.