Computational Audition
Part III: Fundamentals of Computational Audition

DeLiang Wang

Perception and Neurodynamics Lab
Ohio State University
http://www.cse.ohio-state.edu/pnl
Outline of Part III

- Auditory representations
- Multipitch tracking
- Auditory segmentation
- Ideal binary mask as CASA goal
  - Coupling binary masking and automatic speech recognition
Cochleagram: Auditory spectrogram

**Spectrogram**
- Plot of log energy across time and frequency (linear frequency scale)

**Cochleagram**
- Cochlear filtering by the gammatone filterbank (or other models of cochlear filtering), followed by a stage of nonlinear rectification; the latter corresponds to hair cell transduction by either a hair cell model or simple compression operations (log and cube root)
- Quasi-logarithmic frequency scale, and filter bandwidth is frequency-dependent
- Previous work suggests better resilience to noise than spectrogram
Resynthesis from a time-frequency mask

• With a cochleagram, a waveform signal can be resynthesized from a binary or ratio time-frequency (T-F) mask (Weintraub’85; Brown & Cooke’94)

• A T-F mask is used as a matrix of weights on the gammatone filterbank:
  • The output of a gammatone filter at a particular time frame is time-reversed and passed through the filter again. Its response is time-reversed the second time. This is to compensate for across-channel phase shifts
  • The output is either retained or removed for a binary mask, or attenuated for a ratio mask, according to the corresponding value of the mask
  • A raised cosine window is used to window the output
  • Sum the outputs from all filter channels at all time frames. The result is the resynthesized signal

• The resynthesis process can be viewed as an inverse transform from a T-F representation back to the waveform signal
Neural autocorrelation for pitch perception

Licklider (1951)
Correlogram

- Short-term autocorrelation of the output of each frequency channel of the cochleagram
- Peaks in summary correlogram indicate pitch periods (F0)
- A standard model of pitch perception

Correlogram & summary correlogram of a vowel with F0 of 100 Hz
Correlogram response to two sounds

Parts III-IV
Onset and offset detection

• An onset (offset) corresponds to a sudden intensity increase (decrease), which can be detected by taking the time derivative of the intensity

• To reduce intensity fluctuations, Gaussian smoothing (low-pass filtering) is typically applied (as in edge detection for image analysis):

\[ G(t, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \]

• Note that \( (s(t) \ast G(t, \sigma))' = s(t) \ast G'(t, \sigma) \), where \( s(t) \) denotes intensity and

\[ G'(t, \sigma) = \frac{-t}{\sqrt{2\pi\sigma^3}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \]
Onset and offset detection (cont.)

- **Hence onset and offset detection is a three-step procedure**
  - Convolve the intensity $s(t)$ with $G'$ to obtain $O(t)$
  - Identify the peaks and the valleys of $O(t)$
  - Onsets are those peaks above a certain threshold, and offsets are those valleys below a certain threshold
Dichotomy of segmentation and grouping

• Mirroring Bregman’s two-stage conceptual model, a CASA model generally consists of a segmentation stage and a subsequent grouping stage

• Segmentation stage decomposes an acoustic scene into a collection of segments, each of which is a contiguous region in the cochleagram with energy primarily from one source
  • Temporal continuity
  • Cross-channel correlation that encodes correlated responses (fine temporal structure) of adjacent filter channels

• Grouping aggregates segments into streams based on various ASA cues
Cross-channel correlation for segmentation

- Correlogram and cross-channel correlation for a mixture of speech and trill telephone
- Segments generated based on cross-channel correlation and temporal continuity
Sound localization

Figure 5.1  Interaural differences of time and intensity impinging on an ideal spherical head from a distant source. An interaural time delay (ITD) is produced because it takes longer for the signal to reach the more distant ear. An interaural intensity difference (IID) is produced because the head blocks some of the energy that would have reached the far ear, especially at higher frequencies.
Neural cross-correlation

- Cross-correlogram: Cross-correlation (or coincidence) between the left ear signal and the right ear signal
- Strong physiological evidence supporting this neural mechanism for sound localization (more specifically azimuth localization)

Jeffress (1948)
Cross-correlation with one source

- Cross-correlogram and summary cross-correlogram of a vowel with an interaural time difference of 0.25 ms
Azimuth localization of two sources

Cross-correlogram within one frame in response to two speech sources (at 0° and 20°)

Skeleton cross-correlogram sharpens cross-correlogram, making peaks in the azimuth axis more pronounced
Multipitch tracking model of Wu et al.’03

Speech/Interference

Cochlear Filtering

Normalized Correlogram

Channel/Peak Selection

Pitch Tracking Using HMM

Channel Integration

Continuous Pitch Tracks
Channel and peak selection

• Some channels are dominated by interference and provide corrupting information on periodicity. These corrupted channels are excluded from pitch determination in order to reduce noise interference
• Different strategies are used for selecting valid channels in low- and high-frequency ranges
• Further peak selection is performed in high-frequency channels
Channel integration

- How does a frequency channel contribute to a pitch-period hypothesis?
- How to integrate the contributions from different channels?

![Graph showing Ideal Pitch Delay and Peak Delay with Relative Time Lag indicated]

Parts III-IV 18
Relative time lag statistics

- Histogram of relative time lags from natural speech
- Estimated probability distribution of relative time lags as a sum of Laplacian and uniform distributions
Channel integration (cont.)

• Step 1: Taking the product of observation probabilities of all channels in a time frame
• Step 2: Flattening the product probability. The responses of different channels are usually correlated and this step is used to correct the probability overshoot phenomenon
Integrated observation probability distribution
(2 pitches)

Log(Probability)
Last stage of Wu et al. model

Speech/Interference → Cochlear Filtering

Speech/Interference → Normalized Correlogram → Channel/Peak Selection

Pitch Tracking Using HMM → Channel Integration

Continuous Pitch Tracks
Prediction and posterior probabilities

Assuming pitch period $d$ for time frame $t-1$

Observation probabilities for time frame $t$

Prior probabilities for time frame $t$

Posterior probabilities for time frame $t$
Pitch change statistics in consecutive frames

Consistent with the pitch declination phenomenon in natural speech
Hidden Markov model (HMM) for tracking

Viterbi algorithm is used to find the optimal sequence of pitch states
Example: Pitch tracking of two utterances

Wu et al. (2003)

Tolonen & Karjalainen (2000)
Recent developments

- **Jin & Wang (2011)** extend the HMM model to deal with both noisy and reverberant speech
  - In reverberation, the relative time-lag distribution is altered
  - The extended model uses a pitch salience function (summated autocorrelations over selected channels) instead
  - Addresses the question of what the ground-truth pitch should be for reverberant speech
- **Wohlmayr & Pernkopt (2011)** perform multipitch tracking using factorial HMM (FHMM)
  - FHMM provides a means for tracking multiple HMMs
  - Factorial nature leads to efficient inference
  - Allows for speaker-dependent and speaker-independent versions
Auditory segmentation

- The task of segmentation is to decompose an auditory scene into contiguous T-F regions, each of which should contain signal mostly from the same event
  - This definition works for both voiced and unvoiced sounds
- This is equivalent to identifying onsets and offsets of individual T-F regions, which generally correspond to sudden changes of acoustic energy
- One segmentation approach is then based on onset and offset analysis of auditory events (Hu & Wang’07)
  - Onsets and offsets are strong ASA cues for human sound organization
What is an auditory event?

• To define an auditory event, two perceptual effects need to be considered:
  • Audibility
  • Auditory masking
• We define an auditory event as a collection of the audible T-F regions from the same sound source that are stronger than combined intrusions
• Hence the computational goal of segmentation is to produce segments, or contiguous T-F regions, of an auditory event
  • For speech, a segment would correspond to a phone or syllable
Cochleagram and ideal segments

(a) Mixture

(b) Ideal segments for speech

Parts III-IV
Multiscale analysis for auditory segmentation

- From a computational standpoint, auditory segmentation is similar to image segmentation
  - Image segmentation: Finding bounding contours of visual objects
  - Auditory segmentation: Finding onset and offset fronts of segments
- Our onset/offset analysis employs a multiscale analysis, which is commonly used in image segmentation
- Our proposed system performs the following computations:
  - Smoothing
  - Onset/offset detection and matching
  - Multiscale integration
Smoothing

- For each filter channel, the intensity is smoothed over time to reduce the intensity fluctuation with a lowpass filter.
- An event tends to have onset and offset synchrony along the frequency axis. Consequently the intensity is further smoothed over frequency to enhance common onsets and offsets in adjacent frequency channels with a Gaussian kernel.

\[ v(c, t, 0, s_t) = v(c, t, 0, 0) \ast h(s_t) \]
\[ v(c, t, s_c, s_t) = v(c, t, 0, s_t) \ast G(0, s_c) \]

- \( v(c, t, s_c, s_t) \): smoothed intensity at time \( t \) in channel \( c \)
- \( h(s_t) \): a low-pass filter with pass band \([0, 1/s_t]\)
- \( G(0, s_c) \): a Gaussian function with zero mean and standard deviation \( s_c \)
- \((s_c, s_t)\) indicates the degree of smoothing and is referred to as \textit{scale}. The larger \((s_c, s_t)\) is, the smoother \( v(c, t, s_c, s_t) \) is
Smoothed intensity

- Utterance: “That noise problem grows more annoying each day”
- Interference: Crowd noise in a playground
- SNR: 0 dB
- T-F Scale: (a) (0, 0), initial intensity. (b) (2, 1/14). (c) (6, 1/14). (d) (6, 1/4)
Onset/offset detection and matching

- At each scale, onset and offset candidates are detected by identifying peaks and valleys of the first-order time-derivative of $v$
- Detected candidates are combined into onset and offset fronts, which form contours along the frequency axis of the cochleagram
- Individual onset and offset fronts are matched to yield segments
Multiscale integration

• Our system integrates segments generated at different scales iteratively:
  • First, it produces segments at a coarse scale (more smoothing)
  • Then, at a finer scale, it locates more accurate onset and offset positions for these segments. In addition, new segments may be produced

• The advantage of multiscale integration: It analyzes an auditory scene at different levels of detail so as to detect and localize auditory segments of different sizes appropriately
The bounding contours of estimated segments from multiscale analysis. The background is represented by blue:

1) One scale analysis
2) Two-scale analysis
3) Three-scale analysis
4) Four-scale analysis
5) The ideal binary mask
6) The mixture
Goal of auditory scene analysis

• The goal of ASA is to segregate sound mixtures into separate perceptual representations (or auditory streams), each of which corresponds to an acoustic event in the environment (Bregman’90)

• What is the computational goal of ASA?
  • By extrapolation, the computational goal of ASA or the goal of computational auditory scene analysis (CASA) is to develop computational systems that extract individual streams from sound mixtures
Computational-theory analysis of ASA

- To form a stream, a sound must be audible on its own
- The number of streams that can be computed at a time is limited
  - Magical number 4 for simple sounds such as tones and vowels (Cowan’01)?
  - 1, or figure-ground segregation, in noisy environment such as a cocktail party?
- Auditory masking constrains the ASA output
  - Within a critical band a stronger signal masks a weaker one
- ASA result depends on sound types
An obvious goal?

- Extract all underlying sound sources or the target sound source (the gold standard)
  - Implicit in speech enhancement and spatial filtering
  - Segregating all sources is implausible, and probably unrealistic with one or two microphones
Ideal binary mask as CASA goal

- Motivated by above analysis, we have suggested the ideal binary mask as a main goal of CASA (Hu & Wang, 2001; 2004)
- Key idea is to retain parts of a mixture where the target sound is stronger than the acoustic background, and discard the rest
- The definition of the ideal binary mask (IBM)

\[ IBM(t, f) = \begin{cases} 1 & \text{if } SNR(t, f) \geq \theta \\ 0 & \text{otherwise} \end{cases} \]

- \( \theta \): A local SNR criterion (LC) in dB, which is typically chosen to be 0 dB
- Optimality: Under certain conditions the IBM with \( \theta = 0 \) dB is the optimal binary mask in terms of SNR gain
- It does not actually separate the mixture!
IBM illustration

Target

Intrusion

Mixture

Masked Mixture
Properties of the IBM

- **Flexibility**: With the same mixture, the definition leads to different masks depending on what target is.

- **Well-definedness**: The IBM is well-defined no matter how many intrusions are in the scene or how many targets need to be segregated.

- **Consistent with computational-theory analysis of ASA**
  - Audibility and capacity
  - Auditory masking
  - Effects of target and noise types

- **The ideal binary mask provides an excellent front-end for robust automatic speech recognition** (see later)
Subject tests of ideal binary masking

- Recent studies found large speech intelligibility improvements by applying ideal binary masking for normal-hearing (Brungart et al.’06; Li & Loizou’08; Cao et al.’11), and hearing-impaired (Anzalone et al.’06; Wang et al.’09) listeners
  - Improvement for stationary noise is above 7 dB for normal-hearing (NH) listeners, and above 9 dB for hearing-impaired (HI) listeners
  - Improvement for modulated noise is significantly larger than for stationary noise
Test conditions of Wang et al.’09

- **SSN**: Unprocessed monaural mixtures of speech-shaped noise (SSN) and Dantale II sentences (0 dB: 🎤 -10 dB: 🎤)
- **CAFÉ**: Unprocessed monaural mixtures of cafeteria noise (CAFÉ) and Dantale II sentences (0 dB: 🎤 -10 dB: 🎤)
- **SSN-IBM**: IBM applied to SSN (0 dB: 🎤 -10 dB: 🎤 -20 dB: 🎤)
- **CAFÉ-IBM**: IBM applied to CAFÉ (0 dB: 🎤 -10 dB: 🎤 -20 dB: 🎤)

  - Intelligibility results are measured in terms of speech reception threshold (SRT), the required SNR level for 50% intelligibility score
Wang et al.’s results

- 12 NH subjects (10 male and 2 female), and 12 HI subjects (9 male and 3 female)
- SRT means for the 4 conditions for NH listeners: \((-8.2, -10.3, -15.6, -20.7)\)
- SRT means for the 4 conditions for HI listeners: \((-5.6, -3.8, -14.8, -19.4)\)
Speech perception of noise with binary gains

- Wang et al. (2008) found that, when LC is chosen to be the same as the input SNR, nearly perfect intelligibility is obtained when input SNR is \(-\infty\) dB (i.e. the mixture contains noise only with no target speech)
T-F masking and ASR

• Early work found poor ASR performance by directly recognizing the resynthesized signal from a binary mask

• Missing data (feature) approaches have been developed that work with binary masks
  • Marginalization
  • Reconstruction

• Direct masking has been recently shown to be effective
Marginalization

- The aim of ASR is to assign an acoustic vector $X$ to a class $C$ so that the posterior probability $P(C|X)$ is maximized: $P(C|X) \propto P(X|C) P(C)$
- If components of $X$ are unreliable or missing, $P(X|C)$ may not be computed as usual
- Marginalization (Cooke et al.’01) partitions $X$ into reliable parts $X_r$ and unreliable parts $X_u$, and uses marginal distribution $P(X_r|C)$ in recognition
- It provides a bridge between a binary T-F mask generated by CASA and recognition
Reconstruction

- Marginalization is performed in T-F or spectral domain
  - Clean speech recognition accuracy in the cepstral domain is higher
  - Recognition performance drops significantly as vocabulary size increases
- Raj et al. (2004) suggest recognition in cepstral domain after reconstruction of missing T-F units
  - The prior is modeled as a large GMM trained from pooled speech data
  - Missing parts can be reconstructed from the GMM given reliable parts
  - With reconstructed parts, whole frames are converted to the cepstral domain where recognition is performed
Direct masking

• Hartmann et al. (2013) recently found that the conventional wisdom that binary masks cannot be used directly for ASR is a misconception
  • The key ingredient missing in earlier studies is variance normalization of features on a per-utterance basis

• With variance normalization, much simpler direct masking performs at least as well as marginalization and reconstruction
  • For both small and large vocabulary tasks

![Graph showing word accuracy for different conditions](image-url)
Summary of Part III

- **Fundamental representations**
  - Cochleagram, correlogram, onsets/offsets, cross-channel correlation, cross-correlogram
- **Multipitch tracking**
- **Auditory segmentation**
- **The IBM and speech intelligibility**
  - Robust ASR with binary masks
Computational Audition

Part IV: Computational Auditory Scene Analysis

DeLiang Wang

Perception and Neurodynamics Lab
Ohio State University
http://www.cse.ohio-state.edu/pnl
Humans versus machines

Additionally:

- Human word error rate at 0 dB SNR is around 1% as opposed to 40% with noise adaptation
  - At 0 dB
- Roughly speaking. ASR performance in noise is an order of magnitude worse than HR listeners, who are about another order of magnitude worse than NH listeners

Fig. 5. Performance of humans (Ebel and Picone, 1995) and of a high-performance HMM recognizer with noise compensation (Gopinath et al., 1995; Pallett et al., 1995) for Wall Street Journal sentences with additive automobile noise.

Source: Lippmann (1997)
Machine approaches to sound separation

- Speech enhancement
- Spatial filtering (beamforming)
- Computational auditory scene analysis
  - Topic of Part IV
Speech enhancement

- Enhance SNR or speech quality by attenuating interference
  - Spectral subtraction is a standard enhancement technique
  - It works with monaural recordings
  - Limitation: Stationarity and estimation of interference
  - Incapable of improving speech intelligibility in noise
Spatial filtering

Spatial filtering (beamforming): Extract target sound from a specific spatial direction with a sensor array

- Advantage: High fidelity with a large array of microphones
- Capable of improving speech intelligibility in certain configurations
- Not applicable when target and interference are co- or closely-located
- Another limitation: *Configuration stationarity* - What if the target sound switches between different sound sources, or changes its location and orientation?
Outline of Part IV

- Traditional CASA
- Classification-based CASA

Typical architecture of CASA systems
Traditional CASA

- Segregation based on primitive ASA cues
  - Overview of early representative models
  - Tandem algorithm (Hu & Wang’10)
  - Unvoiced speech segregation (Hu & Wang’11)
  - Other developments
Early representative CASA models

- **Weintraub’s 1985 dissertation at Stanford**
  - First systematic CASA model
  - ASA cues explored: pitch and onset (only pitch used later)
  - Uses an HMM model for source organization

- **Cooke’s 1991 dissertation at Sheffield (published as a book in 1993)**
  - Segments as synchrony strands
  - Two grouping stages: Stage 1 based on harmonicity and common AM and Stage 2 based on pitch contours

- **Brown & Cooke’s 1994 *Computer Speech & Language* paper**
  - Computes a collection of auditory map representations
  - Compute segments from auditory maps
  - Group segments to streams by pitch and common onset and offset
Early representative CASA models (cont.)

- **Ellis’s 1996 dissertation at MIT**
  - A prediction-driven model, where prediction encompasses from simple temporal continuity to complex inference based on remembered sound patterns
  - Organization is done on a blackboard architecture that maintains multiple hypotheses
- **Wang & Brown’s 1999 *IEEE Trans. on Neural Networks* paper**
  - An oscillatory correlation model with emphasis on plausible neural substrate
  - Clear separation of segmentation from grouping, where the former is based on cross-channel correlation and temporal continuity
- **Hu & Wang’s 2004 *IEEE Trans. on Neural Networks* paper**
  - Deals with resolved and unresolved harmonics differently, and for unresolved harmonics, grouping is based on AM analysis
  - Computes a target pitch contour
  - Formulates the CASA problem as that of estimating the IBM
A tandem algorithm

- **Pitch detection and voiced speech segregation are closely related problems**
  - Harmonicity is a primary cue for voiced speech segregation
  - Pitch has to be estimated from a mixture. This is a challenging task since interference often corrupts target pitch information
    - Hence it is desirable to remove or attenuate interference before target pitch estimation

- **A “chicken and egg” problem**
  - We want to use target pitch to segregate speech or remove interference
  - We want to remove interference before pitch estimation
Tandem algorithm to address both challenges

• **Key observation**
  - One does not need the entire target signal to estimate target pitch
  - Without perfect target pitch, one can still segregate some target signal

• **Hu & Wang (2010) proposed a tandem algorithm that jointly estimates target pitch and segregates target speech**
  - First perform a rough estimate of target pitch and then use this estimate to segregate target speech
  - With the segregated target, we can generate a better estimate of target pitch, then a better segregation of the target with better pitch information, and so on
  - Good estimation is achieved when the iterative process converges
Tandem algorithm diagram

Auditory features → Initial estimation → Pitch estimation given mask → Mask estimation given pitch → Final estimation

Iterative estimation
Tandem algorithm

- **Initial estimation**
  - Our system estimates up to two pitch periods in each time frame from T-F units with high cross-channel correlation
  - A binary mask is generated for each pitch period
  - Estimated pitch periods are verified and connected into pitch contours according to temporal continuity

- **Iterative estimation**
  - We first estimate the pitch contour from a current binary mask
  - Then we re-estimate the mask for each pitch contour

- **Final estimation**
  - With the results from auditory segmentation, we label T-F units within a segment as a whole instead of labeling them individually
Mask estimation given pitch

• The computational goal is to estimate the IBM

• A simple approach: A T-F unit is labeled as target, if and only if the corresponding response or response envelope has a similar period to the estimated pitch
  • The period of a filter response and response envelope is indicated by the corresponding correlogram

• A more reliable way is to consider the information from a neighborhood of T-F units
  • The algorithm combines the AM cue and the periodicity cue to estimate the probability of a T-F unit being target dominant given a target pitch $\tau$, referred to as $P(\tau)$
Pitch estimation given estimated mask

- The tandem algorithm detects target pitch by integrating information from T-F units that are labeled as the target by the given mask
  - Target pitch is indicated by the peaks in the profile of $P(\tau)$ values within the plausible pitch range
  - At each time frame, the algorithm takes the summation of $P(\tau)$’s across all the T-F units labeled as the target and then identifies the target pitch from the summation

- Pitch tracking with temporal continuity
  - Speech signal exhibits temporal continuity, as does a pitch contour
  - Estimated pitch periods are verified and connected into pitch contours according to temporal continuity
Tandem algorithm illustration

- For a 0-dB mixture of a male and a female utterance
  - Detected pitch contours and their associated simultaneous streams are shown in different colors
  - Ground-truth pitch contours are shown as lines
Pitch determination evaluation

For mixtures of speech with both speech and nonspeech intrusions
Voiced speech segregation evaluation

For a corpus of voiced speech mixed with 10 different intrusions (Cooke’93)
Unvoiced speech segregation

- **Unvoiced speech constitutes about 20-25% of all speech sounds**
  - It carries crucial information for speech intelligibility
- **Unvoiced speech is more difficult to segregate than voiced speech**
  - Voiced speech is highly structured, whereas unvoiced speech lacks harmonicity and is often noise-like
  - Unvoiced speech is usually much weaker than voiced speech and therefore more susceptible to interference
- **K. Hu and Wang (2011) recently performed unvoiced speech segregation from nonspeech interference using periodic signal removal, spectral subtraction, and segment classification**
Periodic signal removal

- With voiced speech segregated, periodic portions of interference are removed by cross-channel correlation

- Advantages of periodic signal removal
  - Periodic signal removal excludes T-F units dominated by voiced speech and periodic interference
  - The removal of periodic components of interference makes the remaining interference more stationary
Spectral subtraction

- Noise is estimated by averaging energy of masked aperiodic T-F units in voiced intervals (stationarity assumption)
- Spectral subtraction

\[ \xi(c, m) = \left( |X(c, m)|^2 - |\hat{N}(c, m)|^2 \right) / |\hat{N}(c, m)|^2 \]

- \( \xi(c, m) \) is the local SNR at channel \( c \) and frame \( m \)
- A T-F unit is labeled as unvoiced speech dominant if and only if

\[ \xi(c, m) \geq 0 \]

- Segments are formed by merging neighboring T-F units
Illustration of spectral subtraction

- Removed voiced speech & periodic signals
- Candidate unvoiced segments
Unvoiced speech grouping

- We analyze energy distribution of unvoiced speech and interference segments with respect to their frequency bounds (shown below).
- Based on this analysis, we group segments with a lower frequency bound above 2 kHz or an upper bound above 6 kHz as unvoiced speech (i.e. simple thresholding).

Open bar: speech; filled bar: noise
Evaluation and comparison

- Large SNR improvements are achieved by simple thresholding
- Substantially outperform representative speech enhancement algorithms: spectral subtraction (SS) and a priori SNR based Wiener filtering (Wiener-as)
Examples of unvoiced speech segregation

<table>
<thead>
<tr>
<th>Noise type</th>
<th>Unprocessed mixture</th>
<th>Voiced speech segregation only</th>
<th>Voiced speech + unvoiced speech segregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fan noise</td>
<td><img src="image" alt="Speaker" /></td>
<td><img src="image" alt="Speaker" /></td>
<td><img src="image" alt="Speaker" /></td>
</tr>
<tr>
<td>Rock music</td>
<td><img src="image" alt="Speaker" /></td>
<td><img src="image" alt="Speaker" /></td>
<td><img src="image" alt="Speaker" /></td>
</tr>
</tbody>
</table>

Input SNR is 0 dB
Other development: Music separation

- **Binary masking can be applied to musical sound separation**
- **A unique problem is overlapping harmonics**
  - Western music favors the twelve-tone equal temperament scale, with musical intervals commonly having pitch relationship very close to simple integer ratio
  - Li & Wang (2009) considered 2-source separation using pitch based labeling
- **Temporal context is used to deal with overlapping harmonics**
  - Based on the observation that the signal from the same source has similar spectral shapes
  - For an overlapped harmonic, labeling is based on the estimated amplitudes of overlapped harmonics from non-overlapped harmonics
Music separation example

Parts III-IV
Traditional CASA: Interim summary

- Main progress occurs in voiced speech segregation
  - Accomplished with minimal assumptions
- Reliable pitch tracking is important for CASA
- Recent work starts to address unvoiced speech segregation
Segregation as classification

- **Monaural speech segregation**
  - A GMM based algorithm to improve intelligibility
  - An SVM based algorithm for speech separation
  - A DNN based large scale training algorithm

- **Binaural speech segregation**
The approach

- Speech intelligibility results reported earlier support the IBM as an appropriate goal of CASA in general, and speech segregation in particular
- Hence the speech segregation problem can be formulated as a binary classification problem
GMM-based classification

- Instead of treating voiced speech and unvoiced speech separately, a simpler approach is to perform classification on noisy speech regardless of voicing.

- A classification model by Kim, Lu, Hu, and Loizou (2009) deals with speech segregation in a speaker and masker dependent way:
  - AM spectrum (AMS) features are used
  - Classification is based on Gaussian mixture models (GMM)
  - Speech intelligibility evaluation is performed with normal-hearing listeners.
Diagram of Kim et al.'s model
Feature extraction and GMM

- Peripheral analysis is done by a 25-channel mel-frequency filter bank
- A 15-dimensional AMS feature vector is extracted within each T-F unit
  - This vector is then concatenated with the delta vectors over time and frequency to form a 45-dimensional feature vector for each unit
- One GMM ($\lambda_1$) models target-dominant T-F units and another GMM models ($\lambda_0$) noise-dominant units
  - Each GMM has 256 Gaussian components
Training and classification

- To improve efficiency, each GMM is subdivided into two models during training: one for relatively high local SNRs and one for relatively low SNRs
- With the 4 trained GMMs, segregation comes down to Bayesian classification with prior probabilities of $P(\lambda_0)$ and $P(\lambda_1)$ estimated from the training data
- The training and test data are mixtures of IEEE sentences and 3 masking noises: babble, factory, and speech-shaped noise
  - Separate GMMs are trained for each speaker (a male and a female) and each masker
A separation example

Target utterance

-5 dB mixture with babble

Estimated mask

Masked mixture
Intelligibility results and demo

UN: unprocessed
IdBM: ideal binary mask
sGMM: trained on a single noise
mGMM: trained on multiple noises

Clean:  0-dB mixture with babble:  Segregated:
Discussion

- GMM classifier achieves a hit rate (active units correctly classified) higher than 80% for most cases while keeps the false-alarm (FA) rate relatively low
  - As expected, mGMM results are worse than sGMM
  - HIT – FA well correlates with intelligibility
- The first monaural separation algorithm to achieve significant speech intelligibility improvements
- Main limitation is speaker and masker dependency
SVM-based classification (Han & Wang’12)

- Feature extraction
  - Pitch: autocorrelation values at detected pitch lags
  - AMS
- Unit labeling using support vector machine (SVM)
- Segmentation
  - Voiced frames: Cross-channel
  - Unvoiced frames: Onset/offset analysis
Support vector machine

- **Train an SVM for each channel (64 channels)**
  - Gaussian kernels are used

- **Issue:**
  - Classification accuracy or HIT-FA (false alarm), which is highly correlated to speech intelligibility (Kim et al.’09)?

- **Re-thresholding technique**
  - Instead of 0, we choose a new threshold to maximize HIT-FA in each channel $c$
  - $\theta_c$ is chosen from a small validation set
Estimated IBM

IBM

SVM labeling

Re-thresholding

Segmentation

Parts III-IV
Classification results

Table 1. Classification results for different noises

<table>
<thead>
<tr>
<th></th>
<th>Speech-shaped</th>
<th>Factory</th>
<th>Babble</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5 dB</td>
<td>0 dB</td>
<td>-5 dB</td>
</tr>
<tr>
<td>HIT</td>
<td>60.14%</td>
<td>69.89%</td>
<td>60.02%</td>
</tr>
<tr>
<td>FA</td>
<td>4.10%</td>
<td>3.89%</td>
<td>8.60%</td>
</tr>
<tr>
<td>HIT-FA</td>
<td>56.04%</td>
<td>66.00%</td>
<td>51.42%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>90.33%</td>
<td>89.60%</td>
<td>86.09%</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIT</td>
<td>59.74%</td>
<td>61.02%</td>
<td>57.39%</td>
</tr>
<tr>
<td>FA</td>
<td>20.70%</td>
<td>16.20%</td>
<td>26.71%</td>
</tr>
<tr>
<td>HIT-FA</td>
<td>39.04%</td>
<td>44.82%</td>
<td>30.68%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>76.25%</td>
<td>78.15%</td>
<td>70.60%</td>
</tr>
</tbody>
</table>

Table 2. Classification results for new noises

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Cocktail-party</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5 dB</td>
<td>0 dB</td>
</tr>
<tr>
<td>HIT</td>
<td>69.44%</td>
<td>72.55%</td>
</tr>
<tr>
<td>FA</td>
<td>7.25%</td>
<td>8.32%</td>
</tr>
<tr>
<td>HIT-FA</td>
<td>62.19%</td>
<td>64.23%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>88.81%</td>
<td>87.00%</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIT</td>
<td>48.32%</td>
<td>56.40%</td>
</tr>
<tr>
<td>FA</td>
<td>25.80%</td>
<td>25.61%</td>
</tr>
<tr>
<td>HIT-FA</td>
<td>22.52%</td>
<td>30.78%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>69.83%</td>
<td>69.99%</td>
</tr>
</tbody>
</table>

Kim et al.

- Compared to Kim et al.'09 system (AMS+GMM), this SVM classifier achieves higher HIT-FA rates, particularly for unseen noises.
Demo

- Female Speech + Factory Noise (0 dB)
  - Noisy speech
  - Proposed
  - IBM

Parts III-IV
Generalization is an extremely important issue for supervised learning.

A straightforward approach to generalization is to train on a large variety of acoustic conditions.

For kernel machines, however, large-scale training is almost intractable due to computational complexity.

We propose to learn more linearly separable and discriminative features from raw data via deep neural networks (DNNs) and then train linear SVM for classification (Wang & Wang’13).

- DNNs can be viewed as hierarchical feature detectors.
Deep neural networks

• Why deep?
  • Automatically learn more abstract features as the number of layers increases
  • More abstract features tend to be more invariant to superficial variations
  • Superior performance in practice if properly trained (e.g., convolutional neural networks)

• However, deep structure is hard to train from random initializations
  • Vanishing gradients: Error derivatives tend to become very small in lower layers, causing overfitting in upper layers

• Hinton et al. (2006) suggest to unsupervisedly pretrain the network using restricted Boltzmann machines (RBMs)
DNN training

- **Unsupervised, layer-by-layer RBM pretraining**
  - Train the first RBM using unlabeled data
  - Fix the first layer weights. Use the resulting hidden activations as new features to train the second RBM
  - Continue until all layers are thus trained

- **Supervised fine-tuning**
  - The weights from RBM pretraining provide the network initialization
  - Use standard backpropagation to fine tune all the weights to a particular task
DNN as subband classifier

- **DNN is used for feature learning from raw acoustic features**
  - Train DNNs in the standard way. After training, take the last hidden layer activations as learned features
- **Train SVMs using the combination of raw and learned features**
- **Linear SVM seems adequate**
  - The weights from the last hidden layer to the output layer essentially define a linear classifier
  - Therefore the learned features are amenable to linear classification
DNN as subband classifier
Large training with DNN

• Training on 200 randomly chosen IEEE utterances from both male and female speakers, mixed with 100 environmental noises (Hu, 2006) at 0 dB (~17 hours long)
• Six million fully dense training samples in each channel, with 64 channels in total
• Evaluated on 20 unseen speakers mixed with 20 unseen noises at 0 dB
DNN-based separation results

- Comparisons with a representative speech enhancement algorithm (Hendriks et al.’10)
- Using clean speech as ground truth, on average about 3 dB SNR improvements
- Using IBM separated speech as ground truth, on average about 5 dB SNR improvements
Sound demos

Speech (carrier: "You will say ...")
mixed with speech-shaped noise at -5 dB

Mixture

Separated

Mixture

Separated

Speech mixed with unseen, daily noises

Cocktail party noise (5 dB)

Mixture

Separated

Destroyer noise (0 dB)

Mixture

Separated
Binaural segregation

- The model by Roman et al. (2003) is probably the first to address sound separation as a binary classification problem
- The IBM provides the training labels for supervised learning
- Segregation amounts to IBM estimation
Model architecture

- **Model architecture**
  - **Binaural Cue Extraction**
  - **Pattern Analysis**
  - **Azimuth Localization**
  - **Resynthesis**

- **Auditory Filterbanks**

- **Target**
- **Noise**

Parts III-IV
Azimuth localization

- Cross-correlogram for ITD detection
- Frequency-dependent nonlinear transformation from the time-delay axis to the azimuth axis
- Locations are identified as peaks in the skeleton cross-correlogram
Binaural cue extraction

• **Interaural time difference (ITD)**
  - Cross-correlation mechanism
  - To resolve the multiple-peak problem at high frequencies, ITD is estimated as the peak in the cross-correlation pattern within a period centering at ITD_{target}

• **Interaural intensity difference (IID or ILD): Ratio of right-ear energy to left-ear energy**

\[
IID_{ij} = 10\log_{10} \frac{\sum_{t} r_{ij}^2(t)}{\sum_{t} l_{ij}^2(t)}
\]
IBM estimation

• For narrowband stimuli, systematic changes of extracted ITD and IID values occur as the relative strength of the original signals changes. This interaction produces characteristic clustering in the joint ITD-IID space.

• The core of the model lies in deriving the statistical relationship between the relative strength and the binaural cues.
3-Source Configuration Example

- Data histograms for one channel (center frequency: 1.5 kHz) from speech sources with target at 0° and two intrusions at -30° and 30° ($R$: relative strength)

- Clustering in the joint ITD-IID space
Binary classification

• Independent supervised learning for different spatial configurations and different frequency bands in the joint ITD-IID feature space

• Define:
  \[
  \begin{align*}
  H_1 & \sim p(x \mid H_1): \text{target dominates} \quad (R_{ij} > 0.5) \\
  H_2 & \sim p(x \mid H_2): \text{interference dominates} \quad (R_{ij} \leq 0.5)
  \end{align*}
  \]

• Decision rule (MAP):
  \[
  M(x) = \begin{cases} 
  1, & \text{if } p(H_1)p(x \mid H_1) > p(H_2)p(x \mid H_2) \\
  0, & \text{else}
  \end{cases}
  \]

• Nonparametric method for the estimation of probability densities \(p(x \mid H_i): \text{Kernel Density Estimation}\)
  • Utterances from the TIMIT corpus are used for training
Example (Target: $0^\circ$, Noise: $30^\circ$)

Target

Noise

Mixture

Ideal Binary Mask

Estimated Mask

Segregated Target

Target 🎤 Noise 🎧 Mixture 🎧 Ideal binary mask 🎧 Result 🎧
# Sound demo

2 sound sources (Target: 0°, Noise: 30°)

<table>
<thead>
<tr>
<th>Noise</th>
<th>Mixture</th>
<th>Segregated target</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Cocktail Party’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siren</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female Speech</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 sound sources (Target: 0°, Noise1: -30°, Noise2: 30°)

<table>
<thead>
<tr>
<th>Noise1</th>
<th>Mixture</th>
<th>Segregated target</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Cocktail Party’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female Speech</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parts III-IV
Speech intelligibility evaluation

- The Bamford-Kowal-Bench sentence database is used that contains short semantically predictable sentences as target.

Two-source ($0^\circ$, $5^\circ$) condition
Interference: babble noise

Three-source ($0^\circ$, $30^\circ$, $-30^\circ$) condition
Interference: male utterance & female utterance
Discussion

• The estimation of the ideal binary mask is based on pattern classification in the joint ITD-IID feature space
  • Training is configuration-specific and frequency-specific
• Estimated masks are very similar to ideal ones
  • High-quality estimation of the ideal binary mask translates to high ASR and speech intelligibility scores
• Binaural segregation employs spatial cues, whereas monaural segregation exploits intrinsic sound characteristics
• A main challenge in binaural segregation is room reverberation
  • See for example, Roman et al. (2006)
Classification-based CASA: Interim summary

• Formulation of the segregation problem as binary classification enables the use of supervised learning

• Studies adopting this classification approach achieve very promising results
  • Particularly for improving speech intelligibility in noise, a long-standing challenge
Summary of Part IV

• The primary feature of CASA is that it is perceptually-motivated
  • Emphasis on perceptual characteristics
  • Emphasis on sound properties
  • In other words, it is content-based analysis

• Advances in traditional and classification-based methods have made CASA a strong approach to real-world sound separation