BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

BINAURAL SPEECH SEGREGATION

by

SATYAVARTA

M. Tech., Indian Institute of Technology, Delhi, 2003
B. Tech., Indian Institute of Technology, Delhi, 2003

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2011

Approved by

First Reader _____

Prof Barbara Shinn–Cunningham, Ph. D.
Professor of Cognitive and Neural Systems
Professor of Biomedial Engineering

Second Reader _____

Prof Steve Colburn, Ph. D.
Professor of Biomedial Engineering
Professor of Cognitive and Neural Systems

Third Reader _____

Prof Dan Bullock, Ph. D.
Professor of Psychology
Professor of Cognitive and Neural Systems

BINAURAL SPEECH SEGREGATION

(Order No. 1998424)

SATYAVARTA SATYAVARTA

Boston University Graduate School of Arts and Sciences, 2008

Major Professor: Prof. Barbara G. Shinn–Cunningham, Professor, Cognitive and Neural Systems and Biomedical Engineering

ABSTRACT

This project aims at developing an unsupervised algorithm motivated by psychophysical evidence for separating out a single speaker of interest from a binaural mixture of voices, building an extensible framework for utilizing multiple available cues. The innovation of the algorithm lies in its transformation of all available cues, including pitch, interaural differences in level and time, etc., into features with associated reliabilities represented as vectors in spectro–temporal space. As motivated by behavioral data, the algorithm follows a two–level process of *segmentation* of syllable length segments followed by *grouping* of segments utterance length time scale. The algorithm is evaluated in three room conditions, on a variety of gender and spatial configurations of talkers. The algorithm achieves significant separation in anechoic conditions. It is shown to be robust to reverberation, and outperforms an overfitted supervised linear discriminant in reverberant situations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Preface

Audition is a fascinating faculty that enables a sweeping range of human experience including communication, entertainment, and surviving and negotiating the surroundings. While humans do not rely on audition for foraging or navigation or hiding from predators as much as some other organisms do, they depend upon it heavily for communicating ideas, from the banal "You want some coffee?" to the profound "I think, therefore I am." The incredibly complex system that enables communication offers a host of phenomena for scientific examination and mimetic engineering, for example:

- How can we determine where a sound is coming from, and how important sound locaclization is for communication?
- What causes hearing loss, and how can we fix it to hear better?
- Is having two ears better than having one?
- Are there brain processes that enhance speech signals that may be mimicked to make machines "hear" better?
- How can we focus on a single voice and ignore the rest when several people are talking simultaneously?

In this dissertation, we investigate this last question, which is known as the "cocktail party phenomenon," and formulate it as a computational problem for which we develop and evaluate a solution based on available psychophysical evidence.

1.2 The Cocktail Party Problem

The cocktail party problem in its most general form refers to extraction of a single speaker of interest from any speech mixture (Cherry, 1953). Progress toward solving this problem advances scientific enquiry into understanding of human audition, and has practical applications to prosthetic devices and beyond. In this document we pursue a solution motivated by psychophysical evidence that seeks to overcome limitations of existing algorithms.

The specific problem addressed in this document is as follows:

> Extract a *single speech stream* of interest from a *two–channel speech mixture* of *two or more talkers* recorded in *any room* using the available *acoustic cues*, not necessarily in real–time

The two-channel input sound mixture is similar to a stimulus received at the two ears of a human subject in a cocktail party situation, and can have background noise and reverberant energy from the echoes in the room. The target speech stream is specified by a characteristic (e.g. the location it is coming from, or the gender of the talker). The problem of *speech separation* is therefore a specific subproblem of the general problem of *source* separation attempts to pick a single sound source – be it music, noise, or speech – from a sound mixture.

The desired output of a segregation algorithm is a single speech stream extracted from the mixture. Motivated by practical applications, the evaluation criteria used in this thesis favor output that is intelligible to human subjects or that yields high recognition scores when fed to automatic speech recognition systems.

The target speech stream is extracted only to the extent that it can be differentiated from the distractors without using linguistic or cognitive cues. Thus, effects such as filling in of words and rhythms during moments of low target to distractor energy ratio are not explained by the algorithm, even though these effects are very important for humans trying to understand speech at a cocktail party (Cepeda et al., 2011).

1.3 Applications

From a practical perspective, a solution to the cocktail party problem can lead to improved hearing aids and communication devices that would benefit from selective enhancement of speech mixture stimuli, as well as pre-processing modules for automatic speech recognition systems.

Hearing–impaired listeners who are otherwise satisfied with their hearing aids report problems in understanding speech in group situations. Automatic speech recognition systems require clean speech from a single talker to work effectively.

From the perspective of scientific advancement, a solution motivated by psychophysical data can provide hypotheses for advancing scientific enquiry into audition. Even a purely computational solution that does not seek to mimic human audition can offer insights into how such a task can be accomplished. Also, the problem has parallels in other human cognitive faculties, like vision, memory, and attention, and any solution will potentially be applicable to similar problems in other fields.

1.4 Philosophy of the current approach

Speech from multiple sources is distinguishable by acoustic cues that can be leveraged statistically for speech stream segregation, a combination of which are exploited by various existing algorithms.

For example, the Blind Source Segregation (BSS) algorithm develops a directional filter which selectively enhances sound arriving from a particular location, eliminating any distractor energy or target energy arriving from any other direction.

Another example, the computational auditory scene analysis algorithm developed in relies upon various auditory cues like common amplitude modulation and mutual similarity in pitch in frequency channels where the target source dominates.

A common problem which these approaches are susceptible to is a degradation in performance in presence of reverberation. Echoes in even small rooms tend to effect their performance much more than that of human listeners.

Speech is a highly redundant signal and in anechoic conditions (absence of echoes), relying on any one acoustic cue may be sufficient to distinguish the target from the distractors. In reverberation, when the cues are degraded, multiple cues may be needed to resolve the target and the distractors. Furthermore, humans may perform well in reverberation because of the way in which they exploit the cues, organizing sound elements hierarchically into increasingly complex sound objects.

In this thesis, a computational auditory scene analysis algorithm (CASA) motivated by psychophysical evidence is developed. The novel elements of this approach which may make it perform robustly in face of reverberation are as follows:

1. Exploit a large set of available acoustic cues in a framework where all cues are treated homogeneously after initial pre-processing. This is especially beneficial in reverberation, where individual cues may be degraded, but jointly may still contain information that can help distinguish the talkers. The homogeneous approach also makes the implementation modular and extensible.

2. Weight each acoustic cue by its associated reliability, e.g. binaural coherence (the height of cross-correlation of two signals) is a measure of how reliably the location of the cross-correlation peak, i.e. the interaural time difference (ITD) is known.

3. Process the input signal at multiple temporal levels, using each cue only at the time scale at which it is a discriminant amongst talkers. The variability of an acoustic feature along time determines the time scale at which it can discriminate one talker from the others. For example, onsets of energy distinguish between sources around the instant they occur, and can only be used at a short time scale, whereas ITD can be used at both short and longer time scales.

1.5 Contribution of The Segmentation-Grouping Algorithmin the context of CASA

The Segmentation-Grouping Algorithm is unique in combining the following ideas, some of which are present separately in extant CASA algorithms:

- Sequential clustering of $t\nu$ pixels along time, and hierarchical clustering at multiple time scales.
- Homogeneous treatment of multiple acoustic features.
- Incorporation of feature confidence with feature values, albeit with limited demonstrated benefit.

Relative merits of the current approach are described in context of the taxonomy of current speech separation algorithms below.

Homogeneous Multi–Feature Algorithms

The algorithm most similar to the segmentation-grouping algorithm is described by Bach and Jordan (2009). Their spectral clustering technique is a data–driven algorithm that learns a distance metric over the space of multiple feature maps, and incorporates multiple time–scales.

A demerit of this approach is that it takes feature confidence into account only for pitch. Furthermore, it performs *simultaneous* clustering over multiple time–scales.

In contrast, the segmentation-grouping algorithm performs sequential clustering at a shorter time–scale, and creates longer–range clusters using information from these temporally local clusters.

However, the segmentation-grouping algorithm assumes a fixed relative importance of features, weighted by feature confidence, whereas the spectral clustering approach is designed to learn the metric.

Implicit clustering by oscillatory networks in Wang and Brown (1999) also combines features homogeneously. However, it does not perform sequential clustering along time, and demonstrates a combination of limited features, namely pitch and ITD.

Single-feature Algorithms

Algorithms performing separation based on spatial cues (ITD and ILD) assign a label either to each time–frame, or to each $t\nu$ pixel. Both global and sequential clustering over the $t\nu$ plane has been described for this classification (Roman et al., 2003; Yilmaz and Rickard, 2004).

Algorithms based on harmonic cues focus on following pitch tracks, and cannot operate over inharmonic regions of the spectrogram, such as regions containing unvoiced speech or unresolved harmonics (Weintraub, 1986; Hu and Wang, 2004).

In contrast to such single-feature algorithms, using multiple cues with corresponding confidence measures makes the segmentation-grouping algorithm performance robust to reverberation.

Multi-feature Combinations with Cue–specific separation technique

The differing ways by which features such as pitch, energy differentials, ITD, and ILD cue separation has led to approaches that retain cue–specific separation techniques (Brown and Cooke, 1994; Ellis, 1996), and attempt ad-hoc combinations of separation achieved by each.

In contrast, The Segmentation-Grouping Algorithm offers a homogeneous framework to accommodate multiple cues, which is extensible to new sources of information.

Feature refinement techniques

Beamforming (Van Veen and Buckley, 1988; Aichner et al., 2002; Baumann et al., 2003) may be viewed as a sophisticated localization technique, and can be used to provide a refined feature to clustering approaches.

Mandel et al. (2010b) refine estimates of ILD and ITD in a supervised model-based approach. In contrast, the segmentation-grouping algorithm is unsupervised and cannot be tailored to a particular room situation; however, it can consume refined feature estimates when available.

Non-CASA techniques

Independent components analysis (Comon, 1994; Bell and Sejnowski, 1995; Torkkola, 1998; Lee et al., 1997; Hyvärinen and Oja, 2000) is a generic signal separation technique for instantaneous mixtures, and has been extended to convolutive mixtures by application in frequency domain, leading to state-of-the-art speech separation systems (Kim et al., 2010).

In contrast, the segmentation-grouping algorithm models speech features explicitly, and does not assume super–Gaussianity of an individual speech signal.

Probabilistic modeling of speech mixtures with latent variables is a supervised technique Smaragdis et al. (2009) that is tuned to specific spatial configuration and relative loudness of specific talkers in specific room conditions. In contrast, The Segmentation-Grouping Algorithmis unsupervised and fully general in application; but with the downside that no specific method is specified to adapt to known mixture conditions.

1.6 Outline of the document

The rest of this document is organized as follows.

Chapter 2 motivates the development of a novel algorithm to speech separation, by presenting the shortcomings faced by current systems that can potentially be offset by incorporating ideas based on psychophysical evidence.

Chapter 3 introduces the framework of the algorithm, and chapter 4 describes the computation and scaling of the various feature maps used in that framework.

Chapter 5 develops the algorithm from a veridical linear separator to an unsupervised sequential clustering algorithm. A purely energy based metric is used for evaluating the algorithm in this development phase. The chapter concludes with results on various mixture configurations, and results on the PASCAL automatic speech recognition task.

Chapter 6 concludes the thesis by summarizing the contribution and the challenges in developing a speech separation system.

Chapter 2

Literature Review

The cocktail party problem has been approached both by scientific studies aimed at investigating the underlying psychological mechanisms, as well as by engineering efforts aimed at designing algorithms to separate speech. This chapter presents a literature survey, and is divided into five sections sections.

Section 2.1 introduces a few basic ideas relevant to acoustic signal processing in general, and specifically to the problem of speech separation, namely the spectro-temporal analysis of sound signals, reconstruction masks for speech separation, the nature of masking, and an introduction to auditory scene analysis.

Section 2.2 presents a summary of various approaches to speech segregation, noting their unique characteristics instead of presenting them in great detail.

Section 2.3 reviews various signal characteristics – inherent (from the origin) or acquired (along the path to the ear/microphone) – that can be aid instantaneous speaker discrimination and therefore source separation. Energy and modulation cues, pitch and harmonicity, and spatial cues are presented with the psychophysical evidence for how human subjects employ them in various tasks, how they are effected by reverberation, and how various current algorithms employ them.

Section 2.4 discusses the issues involved in tackling the problem satisfactorily. The issues identified are the physical problem of reverberation in real rooms, the computational problem of cue combination, and the scientific problem of deciding upon a evaluation metric for separation performance.

Section 2.5 concludes the chapter with a summary of limitations of existing algorithms that motivate the model developed in the rest of this document.

2.1 The Nature of Speech and Masking

2.1.1 Spectro-temporal Representation

The sound received at the two ears is computationally treated as two time series, corresponding to the signal at the left ear s_L and that at the right ear s_R. The cochlea performs decomposition of sound into log-spaced frequency bands of logarithmically increasing bandwidths. Most signal processing techniques operate similarly by band-pass filtering through a filterbank or short-term fourier transform prior to further processing.

A characteristic feature of sound may be computed for each instant in time in various frequency bands to create a spectro-temporal map of that feature analogous to auditory feature maps (Moore, 1987).

Such a feature map is 2-dimensional when the feature is a scalar, and may be multidimensional if the feature being computed is itself a vector. For example, energy contained in the signal at a time-frequency pixel is a scalar, while the short term auto–correlation of a signal at a time instant in a frequency band is a vector.

The simplest spectro-temporal representation of a sound is its spectrogram that represents energy content in each frequency band along time. For a sound mixture, the problem of speech separation is equivalent to computing a spectrogram of a source considering the mixture as a summation of spectrograms of the mixture components.

2.1.2 Reconstruction Masks

The locations on spectro-temporal plane indexed by time instant and frequency band are called *time-frequency ($t\nu$) pixels*. Separating out a single source from a mixture corresponds to determining the proportion of energy at each $t\nu$ pixel of the energy spectrogram originating at that source. Thus, a spectro-temporal map specifying fraction of energy at each $t\nu$ pixel can be overlaid multiplicatively on the mixture spectrogram to obtain a single source.

Some approaches approximate this mask by a *binary mask*, in which a $t\nu$ pixel is labeled as 1 if energy from the target source dominates, and 0 from any of the distractors (or, alternatively, sum of the distractors) dominates. Such a mask provides adequate separation only for mixtures that have little spectro-temporal overlap between the speakers.

2.1.3 Nature of Masking

Psychophysical evidence suggests that masking of the target by distractors is composed of two distinct factors. *Energetic masking* (EM) occurs when the target energy cannot be distinguished from the masker energy in a frequency band at a time instant. Thus, it is the masking that occurs when target energy cannot be detected due to masking from the distracting talkers. *Informational Masking* (IM), on the other hand, occurs when the target energy at an instant can be detected, but the listener finds it difficult to follow the speech stream due to distraction from spectro-temporal components of the masker. EM therefore acts at a local time-scale of a few milli-seconds, while IM operates over longer time scales.

Processes underlying perception have been categorized (Darwin and Carlyon, 1995; Bregman, 1990; Slaney, 1998) into *top-down* and *bottom-up* mechanisms for the purpose of modeling. Bottom-up mechanisms are context-independent, ineluctable data driven processes, leading to creation of the same object from a stimulus regardless of the context in which the stimulus is presented (such as common onset time, pitch and har-

monicity of components, etc.). Top-down mechanisms are context-dependent prediction driven processes, that are perform inference from the stimulus and determine it to be one of a variety of objects depending upon the prevailing context (such as the expected pitch and modulation of speech). While top-down mechanisms create objects in the world depending upon the context, bottom-up mechanisms simplify the stimulus by grouping similar elements together. The human auditory system may be modeled as a system with both these components, possibly with very little modularity.

For such a system, EM operates by disrupting the bottom-up process, while IM disrupts the top-down process by creating multiple possibilities to choose the from.

2.1.4 Auditory Scene Analysis

The organization of an acoustic signal into auditory objects by the human auditory system is termed as auditory scene analysis (ASA) (Bregman, 1990). It provides a framework for discussion of human auditory perception, especially for percepts that cannot be explained by simple statistical regularities.

For example, the spectral content of a four formant complex after all formants onset does not depend upon the time of onset of the individual formants. However, Darwin and Gardner (1986) found a four formant stimulus which is perceived as "ru" when all formants have a common onset and are integrated as a single source, but as "li" when the second formant onsets before the others, and is perceived as a separate source. A computational algorithm operating on the ongoing part of the vowel alone cannot explain this phenomenon.

An understanding of ASA can therefore be helpful in developing computational models for speech separation.

ASA posits that sources are heard as distinct objects, so that any separation occurs not directly at the level of the $t\nu$ pixels of the spectrogram but amongst objects. Thus, speech streams are separated in three stages of *segmentation*, *grouping* and *segregation*. Segmentation of the $t\nu$ plane creates local $t\nu$ segments based on local similarity along any feature, so that in each segment, all pixels belong to the single auditory object. Grouping is the process of relating together various $t\nu$ segments that appear to belong to the same source according to Gestalt principles of proximity, similarity, continuity, closure, and common fate (e.g. onsets, glides, vibrato). This is followed by segregation of the target auditory object created during grouping from the distractors.

Learning and attention play an important role in auditory source separation, but are not discussed in this dissertation.

2.2 Speech Separation Approaches

Existing approaches to model the Cocktail Party phenomenon may be broadly classified (Haykin and Chen, 2005; Vincent et al., 2005b) into statistical pattern classification approaches, and computational auditory scene analysis (CASA) approaches that attempt to separate sources based on principles that govern source segregation by humans.

2.2.1 Statistical Approaches

In these approaches, statistical regularities that characterize a source and distinguish it from others are exploited to separate it from other sources. Thus, the $t\nu$ pixels are labeled as as belonging to the target or the distractor based on their statistical characteristics. The characteristics used are often inspired by the human auditory system, e.g. interaural time difference (ITD), harmonicity, or pitch, which are described in subsequent sections in this chapter.

Statistical approaches are distinguished from CASA by their disregard for auditory object formation.

Parsons (1976) implemented a speech separation algorithm based on pitch differences between the talkers. Instantaneous pitches are computed based on the periodicity of spectral peaks, but are restricted to be within a threshold of the pitches in the previous instant. Thus, pitches are computed and tracked, and the harmonics of one pitch are selected and reconstructed to obtain time-domain signal from a single talker. In the subjective results, the algorithm is described to perform surprisingly well on phonated speech mixtures, and reasonably well for non-phonated speech. The idea of computation of pitch tracks, along with the limitation while dealing with inharmonic energy has carried on to current pitch-based algorithms.

Stubbs and Summerfield (1990) tested two strategies to separate speech using fundamental frequency F_0. The algorithms process the stimulus framewise, and compute cepstrum at each instant. The first algorithm determines $F_0$for both the sources, and filters the cepstrum to eliminate the peak at the masker F_0. The second algorithm evaluated by them is a hybrid approach which uses harmonic selection when the target is voiced, but when it is unvoiced, it uses cepstral filtering to eliminate a voiced masker. They evaluated the algorithms by keyword recognition by normal hearing and hearing-impaired listeners, and found that the cepstral filtering approach yielded better results, and that both the approaches produced lower performance gains for hearing-impaired listeners. They point out that the pitch tracking algorithms could benefit by incorporating continuity of higher formants, and formant amplitude (instead of F_0continuity alone).

Denbigh and Zhao (1992) and Luo and Denbigh (1994) implemented a system that used common harmonicity to create local groups, and combined them using binaural cues and fundamental frequency and continuity cues. Shamsoddini and Denbigh (2001) extended the system to incorporate binaural cues more closely into the determination of fundamental frequency cues. Recent approaches that use auditory cues have been

CASA systems.

Kristjansson et al. (2004) performed high resolution spectral analysis using models of male and female speakers to perform signal separation. Other approaches that use training on mixture components, and use generative models for determining proportion of energy from each talker in each $t\nu$ pixel have been described. Such algorithms are inherently limited in being dependent upon training on the exact circumstances, e.g. room, spatial situation of sources, pitch of talkers, etc. of the mixture to be segregated.

Another approach has been to accurately determine instantaneous spatial cues, and separate speech based on instantaneous direction of the dominant source. The distribution of spatial cues has been used by algorithms as histograms estimated empirically (Yilmaz and Rickard, 2004; Aarabi, 2002). The interaural time difference and interaural level difference cues have also been modeled computationally. Mandel et al. (2007) present an expectation minimization algorithm to estimate model parameters and demonstrate performance on speech separation mixtures.

Beamforming is an array signal-processing approach that implements directional filters that emphasize the sound from a particular direction (Van Veen and Buckley, 1988; Aichner et al., 2002; Baumann et al., 2003). In case of null-steering beamformer, such a directional filter eliminates the sound originating from the estimated direction of the spatially localized masker. For effective source segregation, multiple microphones are required, and the performance degrades rapidly in reverberation.

Blind Source Segregation (BSS) algorithms exploit the statistical independence of the components of the acoustic mixture (Comon, 1994; Bell and Sejnowski, 1995; Torkkola, 1998; Lee et al., 1997; Hyvärinen and Oja, 2000). Such algorithms are based on independent components analysis (ICA) and label $t\nu$ pixels in each frequency band into categories in such a way that the distribution of pixel characteristics within each category is most non-gaussian. This approach requires specification of number of sources in advance and shows degraded performance with non-stationary interference which is a characteristic of reverberant energy. Furthermore, since sources are separated in each frequency band, picking the correct separated stream corresponding to the target at each frequency band presents another challenge.

2.2.2 Computational Auditory Scene Analysis

The most significant characteristic of CASA and CASA-like algorithms is a hierarchical processing of the $t\nu$ pixels in the three steps of segmentation, grouping and segregation.

Weintraub (1986) implemented a CASA system by determining pitch tracks, and grouping harmonic components by common amplitude modulation. He demonstrated the system on mixture of digits recognized by an automatic speech recognition system (ASR), and also on separation of digits from concurrent vowels. Mellinger (1991) created a more elaborate CASA system that employed a larger set of monaural cues for performing segmentation and grouping. He developed the system to handle separation of musical compo-

nents.

A later system by (Brown and Cooke, 1994) implemented a similar strategy focusing on separation of speech mixtures, which used autocorrelation and cross-correlation for creating local $t\nu$ segments, which are then grouped using heuristics based on common onset and common amplitude modulation, allowing the combination of segments only by strong evidence. This model incorporated continuity by frequency glides and multiple instantaneous pitches which had not been used by previous systems.

These systems incorporated various cues but it is not clear how much they benefitted from including them. Also, most of these algorithms attempted to perform separation on monaural mixtures, and did not use spatial cues.

ASA suggests creation of auditory objects, and psychophysical evidence suggests that linguistic and non-linguistic contextual cues retroactively influence perception. This motivates algorithms that, over time, hypothesize objects and track their evolution. Ellis (1996) implemented a prediction driven CASA system, which maintains a "blackboard" of object hypotheses. As the acoustic evidence form sequentially processed time frames accumulates, the hypothesized objects are either augmented so that they become perceptual objects or are eliminated altogether due to lack of evidence.

Neurally motivated algorithms have been proposed to combine multiple cues in a neurally plausible manner. Wang and Brown (1999) have proposed versions of a model in which the network dynamics of oscillating neurons representing various cues arranged in spectro-temporal maps lead to emergence of Gestalt properties described by ASA. This approach circumvents the problem of cue scaling for combination, since all cues are represented in the simulated neurons.

de Cheveigné (1993) designed an approach to extract multiple pitches using sequential comb filtering. Wu et al. (2003) developed a system to track multiple pitches along time, and perform separation based on the fundamental frequency. This is an extension of the Hu and Wang (2002) algorithm which creates local clusters in which one of the mixture sources is dominant. The pitch of the source is then re-estimated based on only the frequency bands included in the cluster instead of all bands at the given instant, thus being CASA-like in creating an acoustic model of the source.

Roman et al. (2003) presented an algorithm for performing separation using binaural cues in a hierarchical manner. Local spectro-temporal clusters are created based on common interaural time difference, and each cluster is then assigned to either the target or the masker stream.

van der Kouwe et al. (2001) directly compared exemplars of the CASA and Blind Source separation approaches and concluded that CASA is a more flexible approach since since it can operate with fewer restrictions. BSS procedures require the mixed sources to be statistically independent, stationary, mixed linearly, and known number of sources and the number of microphones is the same as number of channels. Since reverberation is a convolutive process that violates stationarity and leads to non-linear mixing, this condition is violated in most practical situations.

The following section discusses the cues available for speech separation along with their use in various existing algorithms.

2.3 Cues

Speech is a highly redundant signal, and there are several cues that may characterize a speech stream and distinguish it from other simultaneous speech streams.

2.3.1 Energy

The energy content in each frequency band serves as a characteristic of the sound. Human speech carries most of the energy in bands up to 8 kHz, and is marked by sparsity on the $t\nu$ plane and periods of harmonicity.

Continuity of energy along time in a single band or across proximate bands is a strong grouping cue. The average amplitude of sound from a sound source can also serve as a separation cue, e.g. if the target speech is softer than the distractors but loud enough to be heard, it can be separated from the distractors based on its lower average energy. However, the potential to use energy amplitude as a separation cue is diminished when the masking is mainly energetic.

Computationally, energy map is computed as a short term fourier transform, or by root-mean-squared energy sequentially windowed signal components along log-spaced gammatone frequency bands, called the cochleagram.

Continuity of energy on the spectro-temporal plane is not explicitly modeled by speech separation algorithms. Implicitly, regions of the same source share other common charactersitics that are therefore used instead of energy.

No current models employ the mean energy level of the sources for separation.

Common onset and offset, amplitude modulation and frequency modulation are all characteristics that are visually evident from and can be computationally derived from the energy map.

Modulation

Speech is characterized by its rapidly varying frequency spectrum and amplitude envelope. The amplitude modulation spectrum has maxima corresponding to the rate at which syllables are pronounced, at 3 Hz for frequency bands up to 2 kHz, and shifting to 6 Hz for the 4–8kHz bands (Bronkhorst, 2000).

Common amplitude modulation is a highly effective grouping cue (Bregman, 1990). This can make the

target easier to hear if the modulation of the masker(s) is different from the target, a phenomenon known as co–modulation masking release (Moore, 1999). On the other hand, similar modulation of the target and the masker can interfere with detection of target modulation, causing modulation detection interference (MDI) (Oxenham and Dau, 2001).

Another consequence of amplitude modulation is the purely spectro–temporal effect of hearing out the target during the amplitude minimae of the distractors, known as *dip listening* (Buss et al., 2003). This is a purely spectro–temporal effect operating at the short–term grouping stage in ASA. In particular, it has been found to have little effect as number of distractors increases (Hawley et al., 2004), presumably because dips in multiple distractors rarely align in time. Thus, dip listening appears to allow listeners to make use of short-term peaks in TMR to "glimpse" the target.

Dip listening informs the use of "ideal binary (spectro–temporal) masks" over spectrograms as the objective for speech separation algorithms Brungart et al. (2005). (Rickard et al., 2002) observed that for a typical two talker mixture, $t\nu$ pixels that contained 90% of energy of one source contained 1.1% of the energy from the other source. They formalized the idea of sparsity of source overlap with the notion of approximate W-Disjoint Orthogonality (WDO), which is the normalized amount of overlap of short term fourier transform of two signals s_1 and s_2, which is 0 when the STFTs are disjoint, and 1 when they overlap perfectly.

In contrast to amplitude modulation, common frequency modulation does not appear to be a strong segregation cue (Darwin, 1997), and does not provide any evidence for grouping beyond that provided by instantaneous harmonicity. Many monaural algorithms employ amplitude modulation as a grouping cue (Weintraub, 1986; Hu and Wang, 2004; Brown and Cooke, 1994).

Common onsets and offsets

When a sound source turns on, there is a simultaneous onset of energy in frequency channels dominated by it that serves as a strong grouping cue (Bregman, 1990). The grouping due to common onset is demonstrated in modulation detection interference, which is the inability to discriminate the target modulation from a masker because of common amplitude modulation and simultaneous onset and offset (Oxenham and Dau, 2001). Adaptation in the auditory nerve response emphasizes onsets, preferentially coding spectral changes and onset of new sounds (Darwin and Carlyon, 1995).

Algorithms that use common onset and offset have been described by Hu and Wang (2004), who compute onsets from multi–scale maps, and by Brown and Cooke (1994), who track per–channel energy to detect sudden changes in level. The common aspect of the algorithms is the assumption that the onset marks the beginning of an object which is terminated by a well defined eventual offset.

In reverberation, the offsets are obscured by the reflections of the same source arriving from other surfaces after the source has ceased. The onsets are also obscured by the reflected energy from previously active

sources.

2.3.2 Pitch and Harmonicity

The basic rate at which human vocal tract opens and closes gives human speech a fundamental frequency, or pitch, denoted by F_0. Human pitch ranges from 80Hz to 400Hz, being generally higher for female speakers. Harmonicity is the rate at which

The pitch of a talker averaged over time remains consistent enough that it can distinguish an individual from other talkers. However, the instantaneous pitch varies smoothly with time as an utterance is produced. Figure 4.3 shows the pitch of two talkers. Notice that while the average pitch is discriminable, the instantaneous pitch for the speakers varies with time and may overlap at some instants.

Pitch is available only for instants of time when speech is harmonic, which occurs almost exclusively during production of the voiced vowels, and not during most consonants, unvoiced sounds, fricatives and other unharmonic sounds.

Pitch is a characterisitic of the sound produced by a source. Since sounds from multiple sources are present in a mixture, it no longer makes sense to talk about the pitch of the sound, and we need to introduce the notion of pitches of the component sounds.

Pitch is a perceptually salient feature of the human voice, and a strong grouping cue. Harmonic components of the sound are difficult for human subjects to separate, presumably because a harmonic complex is perceived as a single auditory object. The strong effect of grouping by fundamental frequency acts prior to attentional selection of a speech stream (Summerfield and Culling, 1992), helping create an auditory object that can then be segregated. Common harmonicity of $t\nu$ components favors their perception as a single object even when the channels may each be perceived to be originating from different spatial locations (Darwin, 1997; Darwin and Carlyon, 1995).

This suggests that in an acoustic mixture, harmonicity helps in creation of local groups of $t\nu$ pixels, which are perceived as auditory objects, and the groups belonging to the target source can then be segregated from the rest.

Fundamental frequency cues available to humans is limited by physical limitations. At the auditory periphery, outer hair cells responding to stimuli frequencies beyond 5-6kHz cannot phase lock to those high frequency components, and phase lock to their envelope instead (Moore, 1997). The broadening of tuning curves at higher frequencies leads to decreased resolution at higher frequencies, and allows only about eight harmonics of a periodic complex to be resolved for most pitched sounds (Darwin and Carlyon, 1995).

Pitch detection algorithms that aim to predict human performance take this into account by treating higher frequencies differently from lower frequencies. Specifically, Hu and Wang compute fundamental frequency

at frequencies below 3kHz and compute periodicity of the amplitude envelope at the higher frequencies. On the other hand, Summerfield and Stubbs suggest that by performing separation based on better resolution, they are able to provide separation performance better than can be performed by the human auditory system.

In reverberat, due to temporal smearing of energy in each frequency band, the spectral profile at any instant is contaminated by reflected energy from previous instants, and therefore the instantaneous estimate of pitch is not as strong as it is in anechoic conditions. However, pitch still remains an effective segregation cue.

Pitch and harmonicity are strong cues for separating simultaneous speech streams. Pitch is the most well–utilized cue in most schema–based CASA algorithms (Weintraub, 1986, Hu and Wang, 2004; Brown and Cooke, 1994). Most CASA algorithms use pitch by forming pitch tracks, and grouping $t\nu$ pixels on the spectrogram that correspond to a particular instantaneous pitch. Current methods can extract only the dominant pitch track reliably, limiting the utility of this approach for separating competing speech streams. Multiple pitches may be tracked by recursively subtracting the dominant pitch from the mixture (de Cheveigné, 1998). However, this method cannot handle multiple distractors easily unless the target pitch dominates the mixture, because cancellation of more than a single pitch introduces comb filtering effects and removing components of one pitch also removes components that may be integral multiples of (and therefore components of) another pitch as well.

Timbre

Timbre is the general term for all charactersitics of the sound from a source that distinguish it from other sounds of the same pitch and amplitude. It is the signature of a sound source which includes periodicity, noise, spectral envelope, rise and decay time, change of spectral envelope, frequency modulation, amplitude modulation, prefix at the onsets and suffix at the offsets.

The characteristics used for speaker identification (the mel-frequency cepstral coefficients, and the evolving mel cepstrum) (Campbell Jr, 1997) are examples of timbral attributes. Even though speaker identity is a strong distinguishing cue for sourse separation, it has not been used by any computational algorithm for speech separation.

2.3.3 Spatial Characteristics

Binaural Cues

Binaural hearing makes available three distinct acoustic cues for speech separation.

Interaural time difference (ITD) is the difference in time of arrival of a sound at the two ears, serving to determine the azimuthal location of a source. Human subjects can detect differences in ITDs that are as small as a microsecond. Computational models can determine a related quantity called interaural phase

difference (IPD) by performing frequency band-wise cross-correlation of the binaural channels, and estimate the IPD as the temporal offset at which it peaks. At a particular frequency ν, an IPD θ can be explained by multiple ITDs $k_\nu 2\pi\nu$, where $k_\nu = 1, 2 \ldots$. By combining information across multiple frequency bands, the ITD that can best explain the band-wise IPDs is presumed to be the source ITD.

Binaural coherence is a measure of the strength of correlation of the signals received at the two ears. It is therefore also a measure of the reliability of the ITD estimate computed using cross-correlation. For speech separation, it can be used to discriminate a punctate (or proximal) source from a diffuse (or distant) one. Binaural coherence increases as the signal traverses different paths to reach the two ears, and is decreases as the decorrelation introduced by the two paths increases. Thus, it decreases as the ratio of direct energy from the source to the diffuse reverberant energy decreases. The DTR is used as a measure of source distance, and also as a cue for eliminating reverberant energy.

Interaural intensity difference (IID) is the difference in intensity of a signal arriving at the two ears. Frequencies with wavelength comparable to the head size can bypass the head and consequently do not show big IIDs, whereas smaller frequencies show significant IIDs. Similarity of IIDs in multiple bands signals that they are spectral components of the same source, for $t\nu$ pixels dominated by a single source. The ear facing the source receives the sound at a higher intensity compared to the other ear, and is called the better ear.

The Benefit from Binaural Cues

Binaural and spatial cues help human listeners solve the cocktail party problem at many levels, through acoustic enhancement, low–level auditory grouping, and at the level of streaming.

- Acoustic enhancement of the target occurs if the competing sources are located in different hemifields, so that the target to masker energy ratio (TMR) is higher at the ear facing target than at the other ear. This purely acoustic effect, called the better-ear effect, diminishes when competing sources are distributed in both hemifields (Hawley et al., 2004). This is trivially incorporated in most monaural models by only considering the signal arriving at the better ear.
- Binaural cues are helpful in low–level auditory grouping by allowing near–threshold elements to be detected. A spatially separated tone is easier to detect in noise at a threshold 15 dB lower than the detection threshold in the condition where the tone and noise are co–located (Moore, 1997). This reduction in threshold is called the binaural masking level difference (BMLD), and two low–level binaural effects have been hypothesized to be responsible for it (Colburn and Isabelle, 2001). The first effect operates at low TMR, and enables detection of the target. A dominant masker has consistent ITD and ILD observations, which are disrupted by the target (due to within–channel decorrelation), making the masker observations noisier. This enables detection of the presence of a source besides the

masker. Another effect may come into play at a high TMR, when the interaural parameters observed for the target are consistent (across–channel correlation, which enables the subject to lateralize the target), so that they can cue the co–categorization of $t\nu$ components of the target. While there is scant behavioral evidence supporting this kind of operation (Shinn-Cunningham et al., 2007), many binaural algorithms rely exclusively on this effect (Roman et al., 2003).

- Even when interaural cues are degraded by noise and reverberation, the advantage due to spatial separation persists (Hawley et al., 2004; Shinn-Cunningham et al., 2005), and is robust as the number of interfering sources increases (Peissig and Kollmeier, 1997). Such spatial benefit is explained by Bregman's theory that spatial release from masking involves grouping of sound elements from one direction and segregation of that group from elements of interfering sound in different directions. A number of studies (Hawley et al., 2004; Kidd et al., 2005b) find that spatial unmasking is more robust with multiple speech interferers than with multiple noise interferers. These results give credibility to this argument, since head shadow and binaural unmasking (i.e. factors involved in grouping) are similar for the two conditions, but harmonicity and intelligibility (i.e. the factors involved in streaming and segregation) only are beneficial for speech interferers. In an apparently conflicting result, during a vowel identification task, the subjects are unable to utilize a difference in interaural delays of two vowels for identifying them (Culling and Summerfield, 1995). In any two–stage model, this is explained by the absence of cues at the grouping stage, so that subsequent streaming cannot occur even when interaural cues are available. Use of spatial cues at this level is most likely related to use of perceived location to focus attention on the correct source.

The Mechanism for Binaural Cues

Psychophysical evidence suggests that listeners use ITDs to track an auditory object along time, but cannot use it effectively to group components into an object. Listeners are good at using ITD to track a sound source in space, as opposed to using common roving ITD to group components into a sound source (Darwin, 1997). ITD alone cannot be used to segregate a single talker from simultaneous sound sources in a mixture. Also, grouping by fundamental frequency has been shown to dominate over grouping by location, when spatial cues are noisy or degraded (Darwin, 1997). The role of ITDs in segregating objects has been observed in selective attention tasks where overall performance is generally better when sources are perceived at different locations than when they are perceived at the same location (Shinn-Cunningham et al., 2005).

This evidence supports a model with separate *what* and *where* computations and in which the local object is grouped using some non-spatial cues, and the object is located using interaural and multimodal cues.

Models of speech segregation that utilize spatial cues have focused on grouping together time–frequency chunks having similar interaural characteristics, followed by streaming (Roman et al., 2003; Brown and Cooke, 1994). However, psychophysical studies have consistently shown that binaural cues only weakly

influence auditory grouping, even though they provide a robust and consistent advantage in cocktail party situations. Spatial cues may therefore considered to be primarily effective in perceptual segregation of objects (formed using other cues) rather than in object formation (Shinn-Cunningham, 2005).

The spectro–temporal structure of competing sources is an important factor in comprehension. For multiple interferers, the binaural advantage is only 2-4 dB for noise and speech-modulated noise, but 6-7 dB for speech and time-reversed speech (Hawley et al., 2004; Kidd et al., 2005b). This result suggests that structure in speech and reversed–speech is being used by the auditory system in a way that directly increases the *binaural advantage* in addition to the monaural advantage expected from grouping.

The two ears provide two independent observations for each source, and therefore, the binaural reconstruction of a source from an acoustic mixture can be better than the reconstruction from monaural input.

2.4 Challenges in Speech Separation

2.4.1 Reverberation

Reverberation pulls down listeners' performance in speech separation tasks because of energetic effects (Bronkhorst, 2000) because reverberant energy behaves like noise that masks the target sound. Filling in of amplitude dips of the masker by reverberant energy reduces the gain from masker fluctuations. Interaural level differences (ILD) are also diminished since they are caused by shadowing of energy arriving from the target by the head, and reverberation reduces these ILDs by conveying this energy to the shadowed ear via echoes. ITD cues are degraded because of the spurious ITDs from reflections. Pitch cues are degraded but are relatively resilient.

However, these energetic effects do not lead to dramatic degradation of human performance in speech segregation tasks in all cases. This is because in isolating the target speech stream from a mixture, listeners face not only energetic masking (EM) but also informational masking (IM). EM is the difficulty of detecting the target energy because of the distractor energy, while (IM) is the difficulty in perceptually segregating the (possibly audible) target stream from the competing distractors because of their informational content. For a noise masker, masking is predominantly energetic, but for speech distractors whose words make sense, the masking is informational in addition to being energetic.

For a noise masker that offers purely energetic masking, the benefit of spatial separation decreases from 13 dB in an anechoic chamber to 4 dB in a simulated sound field with reverberation time of 0.25s to 0.4s (Koehnke and Besing, reviewed in (Ebata, 2003)). However, the spatial advantage is maintained to some extent with speech distractors that have energetic masking. Freyman et al. (via (Ebata, 2003)) found that the improvement in signal to noise energy ratio (SNR) threshold due to spatial separation for nonsense speech maskers (offering energetic as well as informational masking) was 14 dB in anechoic condition and

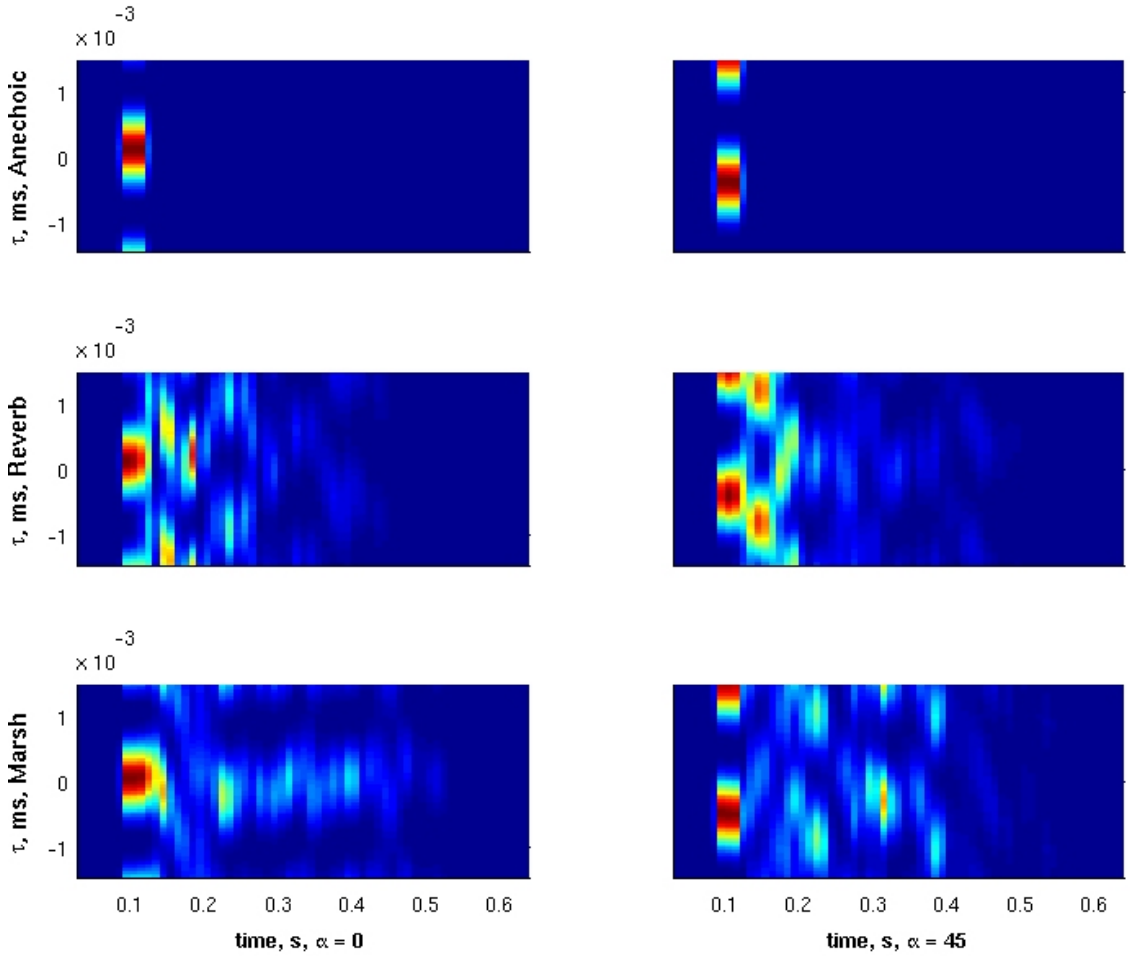


Figure 2.1: Reverberation has two distinct components that impact intelligibility, namely early reflections, and smearing out of energy (Mandel et al., 2010a). The top row shows the left and right channels of what may be considered an anechoic chamber. The second row is a classroom with T_{60} = 0.87 s and the third is a chapel with T_{60} = 1.07 s. The classroom has lower T_{60} but visible early reflections; whereas the chapel does not have prominent early reflections but has reverberation lasting longer.

an appreciable 6–10 dB when a single reflection was introduced. However, with noise maskers (offering energetic masking alone) it changed from 8 dB to 1 dB on introduction of a single reflection.

This spatial benefit for segregating speech mixtures in reverberation is further corroborated by Kidd et al. (2005a), who found that for informational maskers, spatial release from masking was 15–17 dB, and relatively insensitive to reverberation. This suggests that the strengthening of perceptual segregation of sound images due to spatial separation is robust to reverberation, even when short–term binaural cues for separation such as ITD or ILD degrade. Thus, meaningful speech that interferes much more than noise due to informational masking can be streamed and perceptually segregated even in a reverberant environment. For a noise masker that causes energetic masking rather than informational masking, segregation is much simpler and there is no such gain.

2.4.2 Cue Combination

The various cues for distinguishing speech streams provide evidence in disparate units and for different levels of computational abstraction, e.g. ILD is available for each spectro-temporal pixel, while multiple pitches – dominant and others – are available for each pixel, and common onsets and offsets are available for objects rather than for $t\nu$ pixels. Combining the cues to exploit the joint information contained in them, therefore, requires an overarching model that accommodates these abstractions, or interprets the cues in terms of some other abstraction(s).

Most algorithms that utilize multiple-cues are similar to CASA algorithms motivated by psychophysical evidence that suggests a multi-stage process for speech segregation Shinn-Cunningham (2008) as follows: (a) sound must be audible to contribute to intelligibility of a target speech stream; (b) when audible, it must also be segregated from other energy to ensure that its spectro-temporal features can be interpreted; and (c) the listener must then group together the segregated sound segments into streams and select the target speech stream.

Thus, CASA-like algorithms use a subset of available cues at the levels of grouping and streaming in a cue-specific manner, and the essential distinction amongst various algorithms has been the choice of cues used at each level and how those cues are computed and treated. No algorithms has systematically tried to integrate many different possible cues into a single model framework.

Weintraub (1986) uses common pitch and amplitude modulation as cues for grouping and segmentation. Brown and Cooke (1994) use temporal continuity to form segments, which are first merged over frequency bands based on common harmonicity and amplitude modulation, and then streamed based on pitch. Hu and Wang (2004) use a similar algorithm, but use pitch for initial grouping, followed by grouping based on harmonicity and amplitude modulation, and then use a refined pitch estimate for streaming. Roman et al. (2003) utilize spatial cues in a similar scheme but do not use pitch cues at all. Multiple pitch–tracking algorithms (Wu et al., 2003) have been suggested to work with multiple speech sounds, but have not been

employed in any speech separation algorithm implementation yet. Hu and Wang (2007) present an algorithm for monaural speech segregation by utilizing common onsets and offsets. Also, unvoiced speech has been modeled using classification algorithms.

Smaragdis (2001) developed an information theoretic framework for using multiple cues. This instructive approach utilizes Gestalt principles, and is appropriate for musical scene analysis, but does not use binaural cues, and does not explain unmasking in cocktail party situations (i.e., the role of attention in object formation and segregation).

The probabilistic combination of cues suggested by Mandel et al. (2007) suffers from the fact that the training data required for generating empirical probability distributions grows exponentially with the features modeled. Treating the cues as independently distributed trades flexibility for simplicity.

Thus, cue combination has not been adequately tackled in the literature in an extensible way that allows each cue to be treated the same as the others.

2.4.3 Lack of a Consistent Evaluation Metric

Evaluation metrics for speech separation algorithms have been based on sound to noise ratio (SNR) enhancement, speech recognition scores (word error rates), and psychophysical tests comparing subjects' performance on original mixtures versus separated streams (Ellis). However, performance of any algorithm on these three approaches in only loosely correlated, since they reward the slightly different features that they are based upon.

- **SNR Enhancement** For any mixture, the SNR is computed by treating only the target energy as the signal, and the the loss in target energy as well as the inclusion of any distractor energy in the final output as the distractor energy.

  $$\text{SNR} = 10 \log_{10} \frac{\text{target energy included}}{\text{target energy lost} + \text{distractor energy included}}$$

  In the original mixture, the signal is fully included, but the denominator includes all the energy of the target. Thus, this quantity reduces to the target to distractor energy ratio.

  In the output of a functioning speech separation algorithm, the loss of distractor energy raises the SNR by reducing the denominator. Any loss in signal energy is also penalized by the denominator.

  Thus, SNR enhancement can only be computed when the mixture is artificially computed so that the exact target signal content is known.

  Also, this metric does not take into account the speech intelligibility, and does not weigh appropriately the features that may be important for perceptual separation of the sources. For example, a target speech stream is more difficult to listen out in presence of pink noise than when it is masked by

a speech distractor at the same SNR. Thus, distortions may be present at the same SNR but may contribute very differently to the intelligibility of the target.

- **Automatic Speech Recognition Scores**

  Decrease in word error rate (WER) scores produced by automated speech recognition systems on separated streams vs the original mixtures has been used as a metric. However, using such a metric requires co-design of an ASR pre-processor along with the speech separation system to adapt the ASR system to the output of the algorithm.

- **Psychophysical Studies**

  This approach compares the performance of human subjects in recognizing a target in a mixture compared with recognition performance on the target extracted by the separation algorithm. The reconstruction of the target obtained from speech separation algorithms has artefacts and spurious onsets and offsets which may be minor in terms of energy, but wreck its perceptual quality.

  Extracting the target from the masker presents a stimulus akin to the Bregman's letter B demonstration. In a mixture, phonemic restoration occurs allowing filling in of an energetically obscured vowel, but when a phoneme is replaced with silence, subjects find it difficult to fill in. Thus, SNR gains are to some extent offset by the loss in phonemic restoration capacity.

Besides the challenges in meaningfully evaluating speech separation algorithms with WER scores and psychophysical tests, the variety of sound mixture corpora has also led to lack of comparative results of existing algorithms. Artificially spatialized mixtures, instantly mixed speech mixtures, and real room recordings have all been suggested but no standard dataset exists for which results of a number of algorithms exist.

This is partly because algorithms have been specialized to focus on the problem of separating musical instruments, or isolating speech from intrusive noise sources, or separating speech stream mixtures that are either created artificially or recorded in real rooms.

Clear results allowing comparison of algorithm performance on isolating a target speech stream from a mixture are scarce.

2.5 Summary of Limitations of Existing Models

Models based on multiple cues are by no means novel (Weintraub, 1986; Mellinger, 1991; Smaragdis, 2001). However, the model presented in Weintraub (1986) is monaural, based primarily on harmonicity, pitch, and onset and offset detection, and Mellinger (1991) and Smaragdis (2001) are designed for music rather than speech signals and have not been demonstrated on reverberant speech data.

Most existing algorithms employ cues in ways specific to the nature of each cue. Pitch tracks are generally

generated and used for spectral grouping (Weintraub, 1986; Hu and Wang, 2004). Common harmonicity and amplitude modulation are used for local clustering of the $t\nu$ pixels, followed by grouping of local clusters into speech streams based on pitch (Brown and Cooke, 1994). Algorithms relying on spatial cues tend to use interaural level difference (ILD), ITD and other related quantities (interaural phase difference, interaural coherence) to perform classification of the $t\nu$ pixels into target and distractor (Roman et al., 2003; Yilmaz and Rickard, 2004). These algorithms fail when the specific cues that they rely on are degraded in reverberation where one cue may be insufficient to allow segregation.

Most existing algorithms do not consider the reliability of each cue at each $t\nu$ pixel, except through thresholding (ignoring a cue that does not meet the some criterion, such as a threshold on instantaneous energy, or the strength of harmonicity in the case of pitch).

Most algorithms use a single feature or combination of specific features for grouping and streaming. Each feature, after pre–processing, is combined with others in a specific, ad–hoc manner, instead of there being a generalized framework for feature combination that is flexible and takes into account the information reliability of each cue in that pixel.

Furthermore, most algorithms relegate binaural cues to the segmentation process, going against psychological evidence, which suggests spatial cues are primarily used in streaming.

In addition to these general limitations, there are opportunities to improve processing of individual features in a number of ways. Algorithms based on pitch, harmonicity, and other periodic analyses are monaural (Weintraub, 1986; Hu and Wang, 2004; Brown and Cooke, 1994). For humans, spatial benefits are robust in the face of increasing numbers of maskers as well as increasing reverberation. Incorporating the use of spatial cues from binaural inputs should improve these algorithms significantly.

Algorithms using cross–correlation, (e.g. Roman et al., 2003), do not vary the window size of the correlogram analysis based upon the center frequency of the band. There is evidence from psychophysics (e.g. gradual loss of tuning at higher frequencies) that different analysis parameters are appropriate for each frequency. This is used by Hu and Wang (2002) for pitch and modulation, but has not been used for binaural feature computation.

Pitch estimation algorithms work by following pitch tracks, and algorithms to follow multiple pitches have been sought as a solution for separating multiple speech streams. The difficulty of following multiple pitches leads to a limited utility of existing algorithms for speech–on–speech tasks (Wu et al., 2003). However, algorithms that follow a pitch in a given range have not been explored.

In practice, all speech separation algorithms share the limitation of highly reduced performance in reverberant environments. The most exciting opportunity in speech separation is to design an algorithm robust to reverberation.

The following chapters present an algorithm that addresses some of these issues.

Chapter 3

Outline of the segmentation-grouping algorithm

3.1 Motivation

The algorithm to model the cocktail party problem can incorporates the following ideas:

- *Cue Quality*

  The availability of a given cue, and its utility for distinguishing target from masker varies across frequency and time. A measure of cue quality can inform a separation algorithm of the relative extent to which the cue can be relied upon across $t\nu$ pixels. For example, pitch is not available for instants where the speech mixture is inharmonic; and is an unreliable separation cue in high frequencies, and for $t\nu$ pixels where energy from sources is mixed.

  Figure 3.1 compares the pitch map for a two source speech mixture in an anechoic room, to the corresponding map in a real room ($T_{60} = 0.87s$) simulated for the same source mixture using HRIRs. Notice the increased variability across $t\nu$ pixels, and the smearing out of the pitch values from each source along time.

  The mean of harmonicity (along frequency bands) at each instant is plotted as an indicatory of pitch quality, demonstrating the decrease in cue quality during inharmonic time instants in each room, and the decrease due to moving from anechoic chamber to the real room.

  Hence, the cues employed for separation are not equally reliable for all $t\nu$ pixels. Most existing ASA algorithms *select* cues, instead of *weighting* them across the $t\nu$ plane. Algorithms based on pitch are an exception, since they incorporate cue availability by considering only the harmonic segments of speech, but do not incorporate cue reliability in the instants where it is available. A recent algorithm (Mandel and Ellis) exploited the reliability of spatial cues for improved separation.

- *Multiple Cues* A speech signal contains redundant cues to intelligibility. For a signal stripped of this redundancy, such as sinusoidal speech, the intelligibility degrades rapidly on moving from anechoic to reverberant situation, in contrast to harmonic speech whose intelligibility is more robust to reverberation (Poissant et al., 2006). Thus, cue redundancy is valuable in reverberant multi–talker situations.

  A separation model can exploit this cue redundancy by relying on multiple cues. At any instant, only

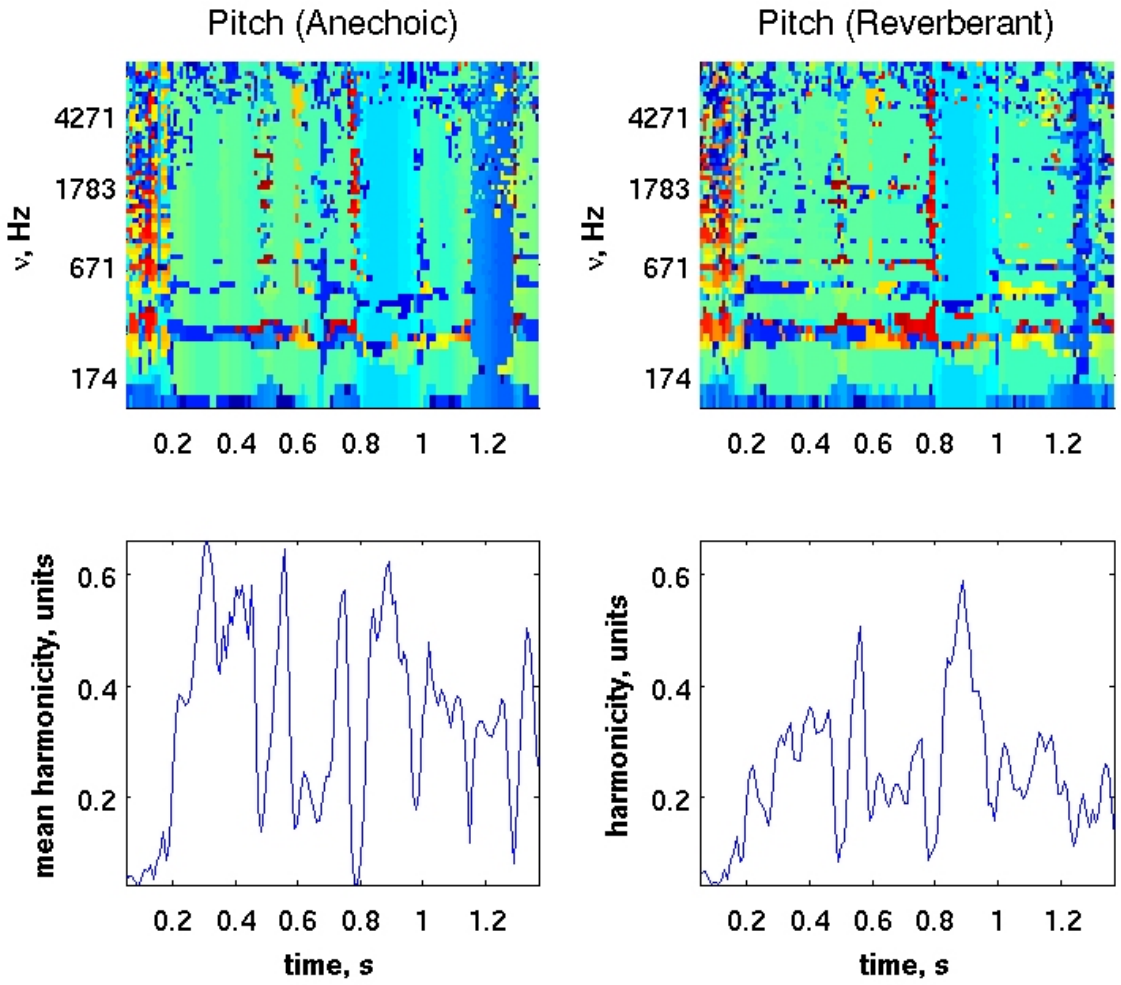


Figure 3.1: Pitch values in anechoic room, compared to a real room. The bottom row plots mean harmonicity in each situation.

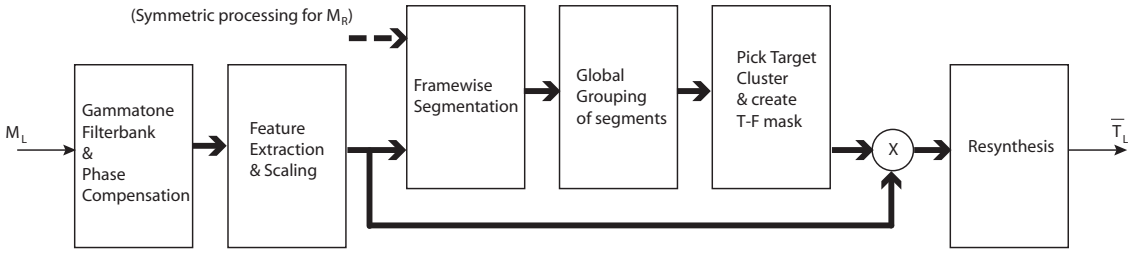


Figure 3.2: The Segmentation-Grouping algorithm

the cues with relatively higher reliability may be used in separation. Hence, in reverberation, when individual cues may not be disparately reliable, the collective information in multiple cues may still enable robust separation.

- *Multiple time scales* Different cues serve as distinguishing characteristics of sources at different time–scales. For example, energy onsets averaged over a long time do not distinguish between the sources, but instantaneously may be used to separate one speaker from another. As another example, there is psychophysical evidence that suggests that human subjects are unable to use spatial cues for creating auditory objects, but do exploit spatial cues for separating auditory objects created based on energetic cues (Shinn-Cunningham, 2005). This indicates that spatial cues are not very helpful for estimating instantaneous spectral content of an auditory object, but help in separating target from masker at a longer time–scale.
- Instead of modeling a top–down process that would be essential for bringing up performance to the level of humans in a reverberant environment, the limits of bottom–up processes that rely only upon instantaneous statistical evidence (and not on any schematic learning) are worth examining.

These charactersitics are embodied in algorithm shown in figure 3.2 as follows:

- Corresponding to each cue, an estimate of its reliability is determined and used as measure of their relative importance. The distance measure in feature space weights the cues proportional to their reliability during local and global clustering.
- Each feature is treated as a dimension along a vector for a given $t\nu$ pixel. This is possible after a transformation, or interpretation, of each cue to a linear metric space. After the transformation, all cues are treated identically, and the information useful for distinguishing sources may be extracted from whatever subset of cues is available.
- Clustering is performed at multiple temporal resolutions: a grouping process occuring in a window of scores of milliseconds, and a streaming process operating over several hundreds of milliseconds. At each tier, cues are selected according to their relative importance in distinguishing sources at the timescale of processing. Most notably, onset cues are used during local clustering, and dropped during streaming.
- No top-down processes are modeled, and therefore, the gains from attentional, cognitive and linguistic factors are not captured by the algorithm.

3.2 Scope

The task of separating speech is intimately interwoven with the problems of linguistic understanding, appreciation of music, and the auditory aspect of multimodal object formation etc. This research is restricted

to designing a speech separation algorithm with the following goals:

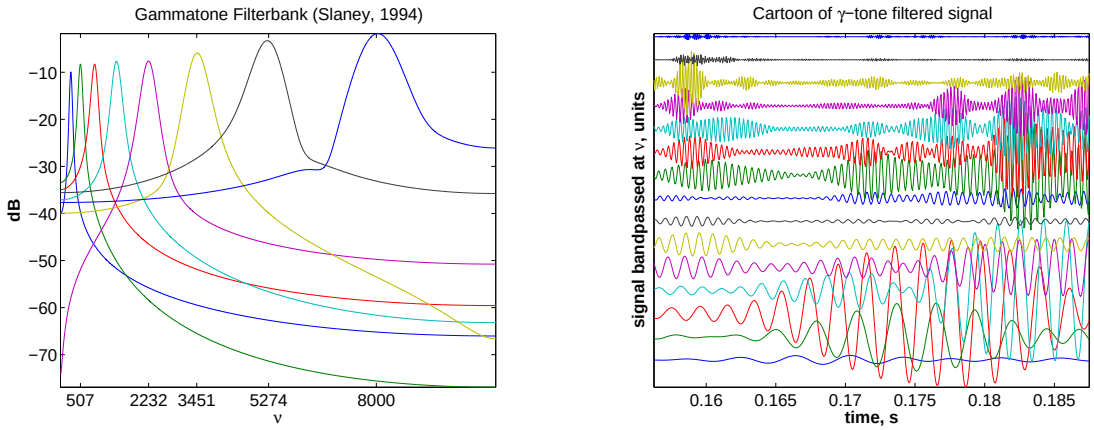
- Separate speech of a single talker from a speech mixture. This is limited to separating speech, and does not aim to reconstruct or infer missing phonetic information obscured by noise, or other acoustic interference.
- Deduce appropriate parameters for the separation algorithm from available psychophysical data for improving separation against evaluation criteria described further in this dissertation. The emphasis lies on improving separation performance, and not on modeling psychophysical data, so that parameters may be derived from means other than available data.
- Evaluate the algorithm on metrics that correlate roughly to speech intelligibility. The metrics chosen have been widely used in the literature, and provide an insight into performance of algorithm against best performance possible under various assumptions. The evaluation does not aim at evaluating algorithm performance in actual speech intelligibility tests.
- Evaluate the algorithm under a variety of stimulus conditions that are varied combinations of gender of the speakers, room conditions (anechoic, mildly reverberant, highly reverberant), speaker separation about the azimuth, and target–to–masker ratio.
- Separate speech to the best extent possible using only bottom–up mechanisms, and this research does not aim at modeling any top–down mechanisms. The process of grouping and segregation of auditory elements exploits purely acoustic commonalities of the acoustic signal without considering contextual or linguistic information. The algorithm does not seek to classify a $t\nu$ pixels into the same source if there is no bottom–up evidence for their co–categorization, even though linguistic evidence may be available. In future work, this approach may be extended to include top–down knowledge about the spectrotemporal structure of speech.
- The algorithm developed here has the primary objective of separating sources using any and all available acoustic information. It is inspired by auditory psychophsics but does not attempt to model any psychophysical data.

3.3 Gammatone Analysis and Resynthesis

Analysis

The time–frequency representation of the mixed signal is a substrate for separation algorithms. The gammatone filterbank (Patterson et al., 2004) approximates the spectral analysis performed by the cochlea, and splits the signal from each channel $S(c,t)$ into an ensemble of signals $S_g(c,t,\nu)$.

The filterbank consists of bandpass filters with log-spaced center frequencies (CF), and bandwidths proportional to respective CFs. To compensate for the phase shift (and align acoustic features across frequencies),



(a) (b)

Figure 3.3: (a) Gammatone filterbank covering the acoustic spectrum audible to humans. (b) Gammatone analysis of a speech signal into a $t\nu$ map. In the processing phase, the value at each time instant is scaled, i.e. suppressed or emphasized to eliminate desired parts of the signal.

Phase	**Parameter**	**Value**
Gammatone Analysis	Frequency limits	200 Hz – 8 KHz
	Number of bands	64
Segmentation	$t\nu$ pixel	10 ms
	Frame width	40 ms
	Frame overlap	20 ms
	Clusters per frame	16
Grouping	Clusters	n = number of sources

Table 3.1: Common parameters for Gammatone analysis/synthesis. Any deviations from these values are specified in the text.

each subband signal $S_g(c, t, \nu)$ is passed through its gammatone filter reversed in time. This is achieved by applying the filter using the MATLABcommand `filtfilt`.

The gammatone filterbank models equivalent rectangular bandwidth (ERB) that enables homogeneous treatment of the subbands in various computations, such as the computation of perceptually significant features (e.g. pitch).

We use the MATLAB implementation from Slaney (1994) as the `ERBFilterBank` function, and performed most of the experiments with 64 logarithmically spaced bands between 100 Hz and 8 kHz (unless specified otherwise).

A more realistic peripheral model, such as the auditory nerve model from Zhang et al. (2001), is not used here since it increases computation time without any clear advantage in feature computation and separation.

FFT spectrogram is not used since the perceptual significance of the frequency spectrum follows a logarith-

mic scale, and hence, a linear scale along frequency overrepresents higher frequency bands.

LPC coefficients that are a staple of automatic speech recognition models are not used here. LPC coefficients summarize the energy content of a single source, that is easier to arrive at after separation of the sources. It is computationally simpler to work with the full $t\nu$ representation for the algorithm developed here.

Computation of a time-frequency mask

The spectrogram of the mixture can be transformed into the spectrogram of the target by applying a target to masker energy ratio mask. Such a per-pixel mask, when derived from veridical knowledge of the target and masker, applied multiplicatively over $t\nu$ spectrogram approximates optimal linear filter for separating sources (Li and Wang, 2009). Even a binary mask created by thresholding the ideal energy ratio mask allows the reconstructed speech to be intelligible in most situations (Wang, 2005), suggesting that the pattern of energy in a source is more important than the actual energy ratio in each pixel (Li and Loizou, 2008).

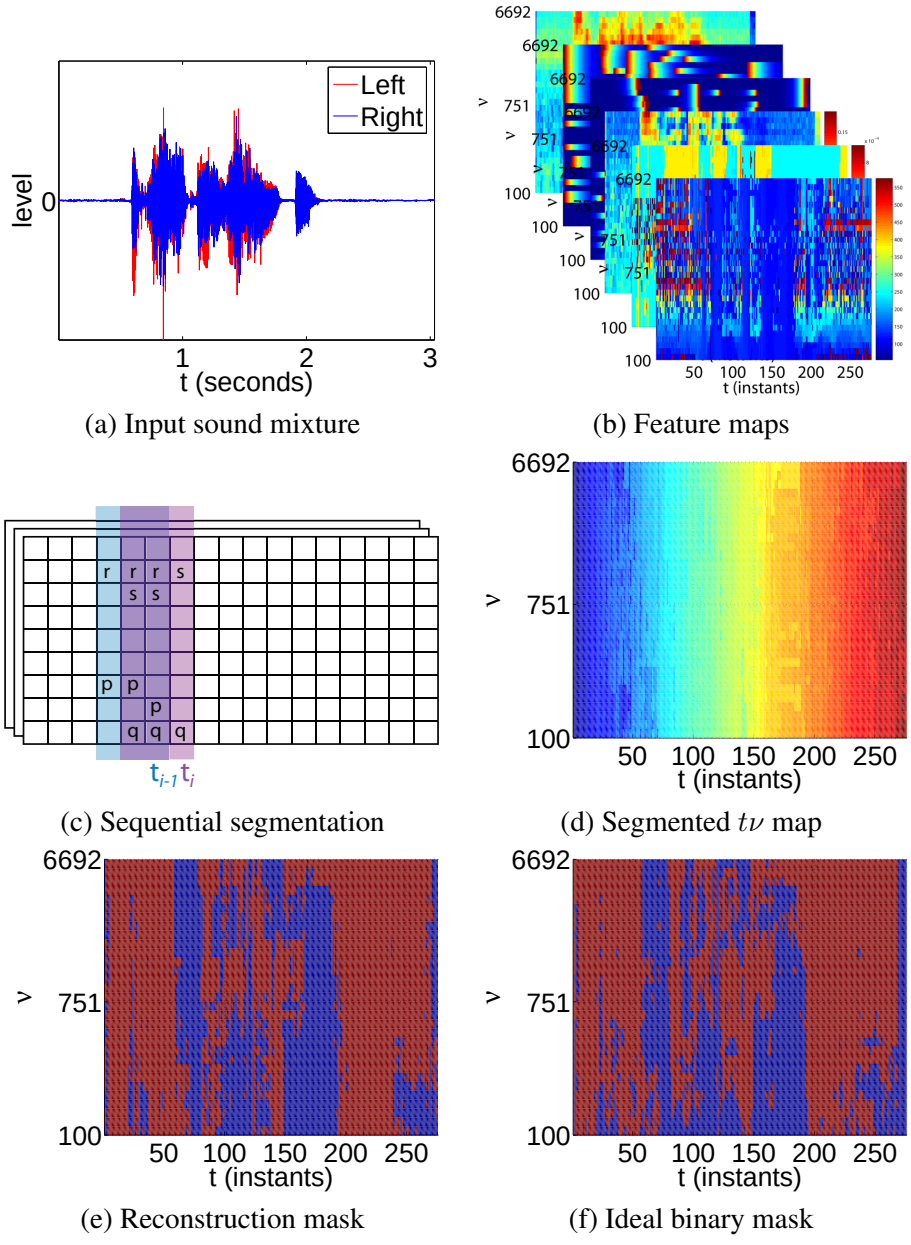
The availability of an ideal binary mask to compare with the reconstruction mask also allows evaluation of speech separation algorithm in the $t\nu$ domain.

The problem of speech sepration is therefore reduced to the problem of estimating the target–to–masker energy ratio mask. An unsupervised algorithm that arrives at this estimate by clustering pixels in multi–dimensional space over multiple time resolutions is described in section 3.4 and forms the core of this thesis.

Resynthesis

In the $t\nu$ representation, signals can be processed in a variety of ways e.g. for dereverberation, noise removal, or as in our case, for eliminating speech from competing talkers. Here, this is achieved by creating a multiplicative weighting mask $w(c,t,\nu)$, so that $S(c,t,\nu)\cdot w(c,t,\nu)$ represents the cochlear decomposition of the output signal.

The output signal is resynthesized as $\hat{S}(c,t)=\sum_{\nu} S_f(c,t,\nu)\cdot w(c,t,\nu)$ (Slaney et al., 1994; Lyon, 1996). Note that the phase compensation step during analysis obviates the need to do so during resynthesis.



(a) Input sound mixture

(b) Feature maps

(c) Sequential segmentation

(d) Segmented $t\nu$ map

(e) Reconstruction mask

(f) Ideal binary mask

Figure 3.4: Sound received by the left and right ears (a) is gammatone filtered to generate cochleagrams for respective channels, which are then processed to obtain a number of feature maps, like pitch, ITD, ILD, energy content, onset and offset maps (b). In the stacked feature maps, each $t\nu$ pixel is a vector, so that a vector distance may be defined between any two $t\nu$ pixels. Frame–wise sequential segmentation is performed (c), and each new grouping is resolved with groupings from previous frames. The segmented $t\nu$ image (d) shows groups of pixels mutually similar at a local scale. Long term similarity amongst groups is exploited to combine $t\nu$ pixels into streams to yield a reconstruction binary mask (e), which can be compared and evaluated relative to the ideal binary reconstruction mask (f).

3.4 Segmentation–Grouping Algorithm

In the $t\nu$ domain, we treat speech separation as an unsupervised classification problem with $t\nu$ pixels to be labeled as either *target* or *masker*. Allowing partial assignment to these categories leads to a fractional reconstruction masks. The final steps are outlined in algorithm listing 1, and described here. For an incremental development of the algorithm, see chapter 5.

Feature map computation

For each gammatone subband of each channel, $S_g(c, t, \nu)$, features are computed at every $10ms$, creating maps in the time–frequency plane. Each $t\nu$ pixel has a multi-dimensional vector of features (listed in table 4.2) associated with it. After creating feature maps in the $t\nu$ plane, pixels are treated homogeneously along the frequency bands.

The acoustic elements from a single source are conjectured to lie in some high–dimensional sub-space of these features. The separability of clusters in this space limits the separation obtained. It is assumed that the available features allow separation of the sources by clustering. This makes some approaches prohibitive, for example, features that may help in separating sources based on spectro–temporal pattern matching cannot be efficiently exploited.

Segmentation

Clustering of $t\nu$ pixels into *segments* is performed at each successive frames of $40ms$ each, overlapping by $20ms$, are considered. Each pixel within each frame is treated as p-dimensional data point to be clustered. Within a frame, the data are clustered into 16 clusters by divisive clustering (algorithm 3), which is more stable than spontaneous clustering by the k-means algorithm. The distance metric in the divisive clustering algorithm takes into account the varying reliability of each dimension. Two approaches for this distance were tried, similar to the treatment of missing data in existing clustering algorithms (Himmelspach and Conrad):

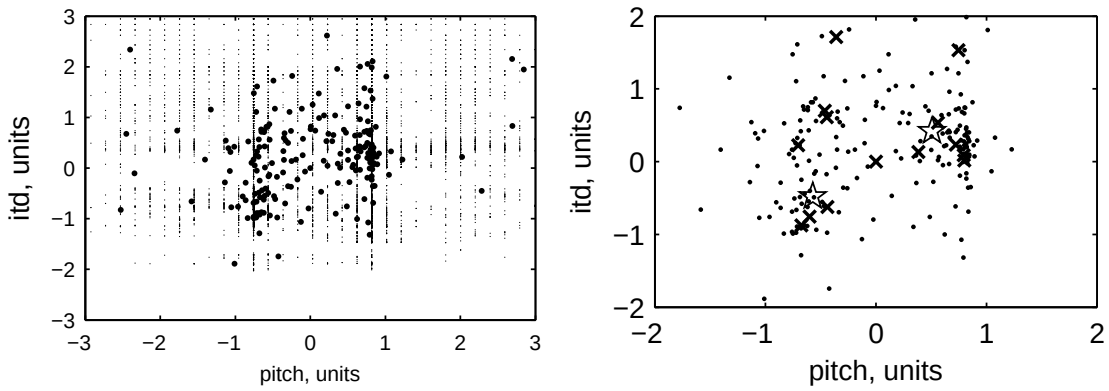
- *Distance Scaling*

$$d_{i,j}^2 = \frac{1}{\sum_f \bar{r}_f} \sum_f \bar{r}_f (x_{i,f} - x_{j,f})^2$$

Algorithm 1: The Segmentation-Grouping Algorithm

Data: Two–channel speech mixture $\mathcal{M}$ sampled at ν_s Hz
Number of talkers n
Result: Reconstructed source signals $\hat{\mathcal{T}}_i, i = 1 \ldots n$

1. *Preprocessing*
   Gammatone filter $\mathcal{M}_L$ and $\mathcal{M}_R$ into b bands to obtain $t\nu$–cochleagram
2. *Feature Maps*
   (a) Compute the p feature maps: Pitch, ILD, ITD, Onset, Offset, Energy, etc. at sampling rate ν_f
   (b) Normalize each feature map to zero mean and unit variance
   (c) Stack feature maps to create a pixel map $\mathbf{F}_{t\nu}$ of p–dimensional feature vectors
3. *Reliability Maps*
   (a) Compute independent reliability maps: Harmonicity, Energy, Coherence, etc.
   (b) Derive dependent reliability maps
   (c) Normalize each reliability map for all values to lie in $[0, 1]$
   (d) Create a map $\mathbf{R}_{t\nu}$ of p-dimensional reliability vectors
4. *Segmentation* using algorithm 2 yields:
   - Local segments $\mathbf{S}$, with segment indices $\mathbf{L}_s$
   - Assignment of pixels to each segment, $\mathbf{P}_s : \mathbf{F}_{t\nu} \rightarrow \mathbf{L}_s$
5. *Grouping* from algorithm 4 yields:
   - Talker models $\mathbf{M}$, with model indices $\mathbf{L}_m$
   - Corresponding pixel assignments $\mathbf{P}_{mi} : \mathbf{F}_{t\nu} \rightarrow [0.0, 1.0]$ for each model $\mathbf{L}_{mi}$
6. *Stream Selection and Reconstruction*
   For each talker model $\mathbf{L}_{mi}$
   (a) Resample $\mathbf{P}_{mi}$ map from the feature sample rate ν_f to signal sample rate ν_s
   (b) Apply resampled map $\mathbf{P}_{mi}$ multiplicatively over $t\nu$–cochleagram to separate the ith hypothesized talker
   (c) Compute time domain signal, $\hat{\mathcal{T}}_i$, from masked cochleagram



(a) Segmentation (b) Coarse and fine grouping

Figure 3.5: Shown here in the feature subspace of pitch vs. ITD, sequential clustering condenses the $t\nu$ pixels into segments (a) that make the inherent clustering apparent. At the streaming stage (b), coarse clusters detect the centers of dominant sources and finer clustering demarcates the segments locally. Selecting one of the dominant sources, marked by the filled asterisk in (c), induces a weighting on finer clusters (circles as targets and squares as distractors), which is allocated to the segments based on their distance from the finer clusters. All $t\nu$ pixels within a segment get the same value in the reconstruction map generated by the ANIMAL .

- *Imputing missing data* The fraction along a dimension that is not reliable is assigned a fixed cost $d > 0$.

$$d_{i,j}^2 = \sum_f (\bar{r}_f(x_{i,f} - x_{j,f}) + (1 - \bar{r}_f) \cdot d)^2$$

The k-means algorithm is equivalent to gaussian mixture modeling, with the assumptions of spherical variance and hard assignment to classes.

Segments created in successive frames are merged if their mutual distance is less than a threshold (set at $0.1p$, which is a 0.1 neighborhood along each dimension).

At this time-scale, several cues are informative, including common onset and offset of energy, common pitch, harmonicity, amplitude modulation, temporal continuity, frequency variation, etc.

At the end of this sequential clustering, the data get summarized into N local clusters, which are much fewer than the original $t\nu$ pixels (figure 3.5a). In subsequent stages of the algorithm, each local cluster is treated as a unit, so that all points belonging to a single cluster are assigned the same weight in the final mask.

Grouping

Grouping of local segments into sound streams, and target stream selection occurs over the time scale of more than 500–1000 ms. Here, we perform grouping over the entire duration of roughly 3s long mixtures.

Algorithm 2: Sequential Segmentation

Data: Maps of features, $\mathbf{F}_{t\nu}$, and reliabilities, $\mathbf{R}_{t\nu}$
Result: Segment centers $\mathbf{S}$ with labels $\mathbf{L}_s$
Segment assignment of pixels $\mathbf{P}_s : \mathbf{F}_{t\nu} \rightarrow \mathbf{L}_s$

Define frame $F_{(t,\, t+l)}$ as pixels lying in time interval $(t,\ t+l)$ from all frequency bands.

1. *Initialization*
   Set $t = 0$, so that the first frame is $F_{(0,\, l)}$
2. *Clustering*
   - Cluster pixels in frame $F_{(t,\, t+l)}$ into k clusters $\mathbf{S}_t$
   - Obtain pixel assignment $\mathbf{P}_{s(}t,\ t+l) : \mathbf{F}_{t\nu,(t,\, t+l)} \rightarrow \mathbf{S}_t$
3. *Reconciliation*
   Resolve groups in $\mathbf{S}_t$ with groupings from previous frames $\mathbf{S}_{0 \ldots t-}$
   (a) For each $\mathbf{S}_{ti}$ in $\mathbf{S}_t$, determine group in $\mathbf{S}_{0 \ldots t-}$ closest to $\mathbf{S}_{ti}$
   (b) If $d(c_i, c_j) < \delta$, merge $\mathbf{S}_{ti}$ into $\mathbf{S}_j$
      - Re–estimate center $\mathbf{S}_j$ to assimilate member pixels from $\mathbf{S}_{ti}$
      - Relabel mappings in $\mathbf{P}_s$ from $\mathbf{S}_{ti}$ to $\mathbf{S}_j$
      - Remove $\mathbf{S}_{ti}$ from $\mathbf{S}_t$
4. *Frame shift*
   $t \Leftarrow t + a$
5. *Termination*
   Iterate over steps 2–4 until t exceeds signal length.

For the results presented, we implemented grouping by clustering the N segments into the same number of groups as the number of sources in the mixture. The clustering algorithm is the same as the one used for segmentation, except that it operates on pixel clusters at this stage.

One of the groups is selected as the target, and assigned unit weight, and the remaining groups are labeled as maskers with zero weight. These weights are them propagated to the N segments, in proportion to their relative distance from the target and the maskers to arrive at the reconstruction mask for the target.

The target selection problem, i.e. selecting one of the n final groups as the target, is bypassed in this results presented here, since it can be tuned to the real-world application. Here, for an n talker mixture, n different masks are created and compared with the veridical mask, and the best performing one is assumed to be marked as the target by the algorithm.

At this longer time-scale, only spatial cues (ITD and ILD) are reliable for stream segregation, and pitch is usable for talkers with marked differences in pitch. In general, greater overlap of pitch is possible among different sources due to local excursions of one speaker into the pitch range of another speaker over time.

Algorithm 3: Divisive K–Means Clustering with Reliabilities

Data: p–dimensional feature vectors $\mathbf{F}$
p–dimensional reliability vectors $\mathbf{F}$
Pixel weights $\mathbf{W}$
Desired number of clusters k

Result: Set of cluster centers $\mathbf{S}$
Vector mapping $\mathbf{P} : \mathbf{F} \rightarrow \mathbf{S}$

1. *Initialization*
   Assign $\tilde{\mathbf{F}} = \mathbf{F}, \tilde{\mathbf{R}} = \mathbf{R}, \mathbf{S} = \emptyset$
2. *Two–way split*
   (a) Randomly pick 2 pixels as centers $\mathbf{C}$
   (b) *Expectation*
      i. Compute distances for each vector $\tilde{\mathbf{F}}_i \in \tilde{\mathbf{F}}$ from each center $\mathbf{C}_j \in \mathbf{C}$
         $d(\tilde{\mathbf{F}}_i, \mathbf{C}_j) = \sqrt{\sum_{m=1}^{p} \tilde{\tilde{\mathbf{R}}}_{i,j;m} (\tilde{\mathbf{F}}_{im} - \mathbf{C}_{jm})^2}$
         where $\tilde{\tilde{\mathbf{R}}}_{i,j}$ is the dimension–wise mean of reliability vectors corresponding to $\tilde{\mathbf{F}}_i$ and $\mathbf{C}_j$.
      ii. Map $\tilde{\mathbf{F}}_i$ to cluster in $\mathbf{C}$ that it lies closest to. $\mathbf{P}_i = \arg\min_{\mathbf{C}_j \in \mathbf{C}} d(\tilde{\mathbf{F}}_i, \mathbf{C}_j)$
   (c) *Maximization*
      Compute centers
      $\mathbf{C}_k = \sum_{i \forall P_i = k} \mathrm{W}_i \langle \tilde{\mathbf{R}}_i, \tilde{\mathbf{F}}_i \rangle$
      $\tilde{\mathbf{R}}_k = \sum_{i \forall P_i = k} \mathrm{W}_i \langle \tilde{\mathbf{R}}_i, \tilde{\mathbf{R}}_i \rangle$
   (d) Iterate steps 2b through 2c until labels converge.
3. *Incorporation in* $\mathbf{S}$
   (a) Update labels of pixels to point to the cluster they belong to $\mathbf{P} : \tilde{\mathbf{F}} \rightarrow \mathbf{C}$
   (b) Add new clusters into the global grouping $\mathbf{S} \Leftarrow \mathbf{S} \cup \mathbf{C}$
4. *Candidate cluster for split*
   (a) Identify cluster with maximum variance in $\mathbf{S}$ as $\tilde{\mathbf{F}}$ and the corresponding reliabilities as $\tilde{\mathbf{R}}$
   (b) If $|\mathbf{S}| < k$, remove $\tilde{\mathbf{F}}$ from $\mathbf{S}$
5. *Termination*
   Iterate steps 2–4 until $|\mathbf{S}| = k$

Still, such information might expected to be beneficial in reverberation, but the proposed algorithm was not able to exploit it (figure 5.18).

Algorithm 4: n-ary Grouping of Local Segments

Data: Segments **S**
Grouping Dimensions
Number of speaker n
Result: Talker models **M**
segment attribution maps $A_{ij} : \mathbf{S} \rightarrow (0.0, 1.0)$
n pixel attribution maps $A_{Fk} : \mathbf{F} \rightarrow (0.0, 1.0)$

1. *Clustering over* **S***, or model creation*
   Create n clusters over **S**, called groups **M**, that approximate long term speech segments
2. *Attribution creation*
   Determine attribution of each segment $\mathbf{S}_i \in \mathbf{S}$ to each group $\mathbf{M}_j \in \mathbf{M}$ based on proximity, $A_{ij} : \mathbf{M}_i \rightarrow (0.0, 1.0)$
   $$\text{Lkhd}(\mathbf{M}_j \,|\, \mathbf{S}_i) = \frac{\Pr(\mathbf{S}_i \,|\, \mathbf{M}_j) \cdot \Pr(\mathbf{M}_j)}{\sum_j \Pr(\mathbf{S}_i \,|\, \mathbf{M}_j) \cdot \Pr(\mathbf{M}_j)}$$
3. *Propagate attribution from segments to pixels*
   Propagate attributions of each segment in **S** to member pixels in **F**, $A_{Fk} : \mathbf{F} \rightarrow (0.0, 1.0)$

return M, A_{ij} *and* A_{Fk}

3.5 Evaluation Metrics

The most relevant measures of separation quality are task-based. For example, in cases where speech separation is applied in automatic speech recognition (and related tasks, like diarization, speaker identification and verification) or prosthetics, the gain achieved on the final task is the most effective evaluation metric.

In the $t\nu$ domain, the intelligibility of the separated signal is approximated by energy based metrics that summarize the spectrogram. These metrics listed below differ in how they combine the per pixel target energy restored compared to the per-pixel distortion present in the masked spectrogram. When the veridical separated signals are available, these methods can be used to compare the intelligibilty scores of the target in the original mixture with the separated signal.

- *Percentage of correctly classified pixels* computed over the $t\nu$ plane. This measure assumes sparsity of the mixed signals, so that $S_{t\nu} \approx max(\S_{\mathcal{T},tf}, \S_{\mathcal{O},tf})$. The simplicity comes with the limitation that the penalty for misclassifying a high energy pixel that leads to a drop in intelligibility, is the same as that for misclassifying a low–energy pixel that may not impact intelligibility at all.
- *Signal to distortion ratio*: The enhancement in sound to distortion ratio (SDR) in going from the

mixture spectrogram to the masked spectrogram is used as the primary metric for comparison. This metric is described in Vincent et al. (2005a) and we adapted it for computation from $t\nu$ masks as follows:

$$SDR_{enh} \quad = \quad 10\log_{10}\frac{\mathcal{E}_{\tilde{s}}}{|\mathcal{E}_{\tilde{s}} - \mathcal{E}_{\hat{s}}|} - 10\log_{10}\frac{\mathcal{E}_{\tilde{s}}}{|\mathcal{E}_{\tilde{s}} - \mathcal{E}_{m}|}$$

Here, $\mathcal{E}_{\tilde{s}}$ is the energy of the source reconstructed from the veridical fractional mask, $\mathcal{E}_m$ is the energy in the mixture, $\mathcal{E}_{\hat{s}}$ is the energy in the reconstructed source, and $(x)_+$ denotes the maximum of x and 0. With this metric, the A1 mask shows 0 dB SDR enhancement in all conditions, the VBM shows the limit of performance of all binary reconstruction masks, and the perfect mask produces ∞ dB SDR enhancement.

- Because of the non–stationary nature of speech, the SDR computed over short time–frames instead of over the whole $t\nu$ spectrogram is a better measure of instantaneous quality. The framewise computation of an SDR–like metric which is averaged across time have been used in various standard intelligibility measures (SII, PESQ, adjusted SII) and also used elsewhere in the literature (Brown and Cooke, 1994).

$$SDR_{adj} = \langle SDR(t) \rangle$$

We undertake evaluation of the output on an Automatic Speech Recognition task publicly available under PASCAL (2007).

Perceptual evaluation may be performed by comparing human listening performance on the reconstructed signal and the mixed signal to quantify the enhancement due to the separation algorithm under various obstructed listening tasks. Perceptual evaluation was not pursued here due to less than impressive separation results in energy based evaluation and the Automatic Speech Recognition task.

Chapter 4

Features

This chapter discusses feature vectors computed per $t\nu$ pixel input to the separation algorithm, enumerated in table 4.1. For a $t\nu$ pixel, the feature vector is a tuple of preprocessed cues computed over a time–window around that instant, for that frequency band.

The considerations for processing the feature and reliability maps are described in section 4.1. Computation of individual feature maps and associated reliability maps is described in subsequent sections.

4.1 Common Considerations in Feature Preprocessing

4.1.1 Scaling and Linearity

Algorithm 3 measures the distance measure between vectors $\tilde{\mathbf{F}}_i$ and $\mathbf{C}_j$ by a modified Euclidean norm:

$$d(\mathbf{F}_i, \mathbf{C}_j) = \sqrt{\sum_{m=1}^{p} \bar{\mathbf{R}}_{i,j;m}(\mathbf{F}_{im} - \mathbf{C}_{jm})^2}$$

This suggests the following constraints on feature representation:

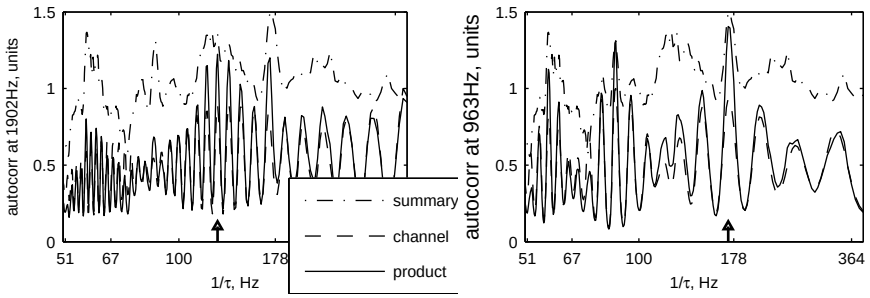
- Linearity: A unit change in feature value should correspond to a unit change in its percept.

  For example, the percept of energy intensity follows Weber's Law, so that the logarithm of actual intensity corresponds to the linear scale of perceived energy in a signal.

  As another example, in case of ITD, the minimum angle discriminable to human subjects is finer when the source is straight ahead than when it is at the side (Saberi and Perrott, 1990). A unit change in ITD feature value does not correspond to a unit change in percent throught the range of values assumed without an appropriate correction, be it non-uniform offset or scaling.

  Feature scaling can be calibrated against psychophysical data, but we do not attempt it, and employ only gross approximations here, e.g. logarithm of energy, actual value of ITD and pitch, etc.

- Comparability along frequency: A unit change in feature value in *any frequency band* must produce equal change in percept.



(a) (b)

Figure 4.1: (a) Autocorrelation of single band around a time instant, showing a spurious peak suppressed by the summary autocorrelation. (b) Two sources have comparable influence on a pixel, and forced to choose one.

- Comparability across features: A unit change in value of *any feature* within a frame must produce equivalent change in percept.

  For linear features, this implies mutual calibration amongst the features. In general, this scaling is achieved by normalizing each feature by its sample variance.

  The relative scaling of features also depends on the noise in the cue, the residual information conveyed by the cue, and its reliability.

  - *Noise* Consider the feature–wise distance between two points along two features, d_{noisy} and d_{clean}. Clearly, the net distance should emphasize the feature with less noise. This is achieved by normalizing each feature by its sample variance.
  - *Information* Clustering is notoriously prone to duplicate information as well as for the inability to reject noise. If a cue does not provide any additional information besides the features already used, it should be scaled down to zero.

    Consider the extreme case of including two copies of a feature in the feature vector; and assume that it is independent of all the other features. The first copy provides information for separation. Including it the second time biases separation in favor of information provided by this cue against that provided by the others.

    In the current algorithm, the exact approximation of mutual information is not performed. All features considered are treated as independent except in situations such as when pitch maps for both left and right channels are used. In such cases, each feature is scaled by $\sqrt{0.5}$ to reflect their dependence.
  - *Reliability* A cue should be relatively scaled down if it is less reliable than other available cues. This factor is represented in the distance metric by the average reliability multiplier.

There are a few more notable implicit assumptions:

- *Each feature must be defined for each $t\nu$ pixel.* Features defined per pixel meet this condition trivially,

Feature	Description
Pitch	Location of autocorrelation peaks
ITD	Location of cross–correlation peaks
ILD	Log of ratio of energy at the two ears
Energy	Log of energy content
Common onsets	Positive part of energy differentials
Common offsets	Negative part of energy differentials
Frequency proximity	Log of center frequency of the band
Harmonicity	Height of the autocorrelation peak
Coherence	Height of the cross–correlation peak

Table 4.1: Features discussed in this chapter

e.g. $\text{ILD}_{t\nu} = 10\log(E_{r,t\nu}) - 10\log(E_{r,t\nu})$. Special processing is performed for features defined contextually.

For example, pitch is defined per *talker*, and at an instant t, k values of pitch are available, one for each of the k active talkers. One of the k values is then assigned to each frequency band, corresponding to the source dominant in that band.

For features assuming a single value per frequency band (or per time instant), the values are replicated across time (or frequencies). For example, to cluster neighboring frequencies together, the center frequency of each band itself is treated as a feature, $\mathbb{F}_{t\nu} = \nu$, and this is constant over all values of t.

- *Feature values are scalar* This prevents, for example, the ability to assign multiple pitches to a $t\nu$ pixel. When multiple pitches influence a pixel almost equally, the map is forced to pick one (figure 4.1), thereby ignoring any per-pixel information about proportion of energy from multiple sources.

  Thus, the algorithm effectively assumes that each pixel is dominated by a single source, and therefore assumes that the feature estimates for any $t\nu$ pixel are estimates, albeit noisy, of a distinct source. Hence, pixels dominated by multiple sources should be weighted down when estimating feature estimates of separated sources.

4.1.2 Feature Reliabilities

The reliability of a feature is a measure of residual information gained from it. The relevance of a feature to separation is not known *a priori*, and in practice, the confidence in feature value is instead used. For example, if two talkers are both located straight ahead, the ITD for both will point to $0s$ delay with high confidence, but the information contained in the cue is negligible (for purposes of our algorithm).

Feature values participate in two distinct computations where reliability can modulate their contribution:

- *Estimating cluster centers:* Each cluster represents a hypothetical source, and the cluster center is the maximum likelihood estimate of the source characteristic. Since each component of the feature vector

is computed as a weighted sum independent of the others, the relative scaling of reliabilities across features does not influence this computation.

$$\begin{aligned}\mathbf{C}_k &= \sum_{i \forall P_i = k} \mathrm{W}_i \langle \tilde{\mathbf{R}}_i, \tilde{\mathbf{F}}_i \rangle \\ \tilde{\mathbf{R}}_k &= \sum_{i \forall P_i = k} \mathrm{W}_i \langle \tilde{\mathbf{R}}_i, \tilde{\mathbf{R}}_i \rangle\end{aligned}$$

The averaging treats reliability as a weighting to reduce the relative contribution of *less reliable pixels*.

A good reliability measure for a given feature is one that balances contribution from feature values at various instants to estimate the actual value of the feature for a single source.

- *Assigning a pixel to a cluster:* The relative distance of a pixel from cluster centers is a proxy for the fraction of energy contributed by each source to that pixel. The actual assignment may be all–or–none, so that entire energy of the pixel belongs to the source it is closest to; or it may be fractional, so that partial energy is attributed to each cluster in proportion to its proximity.

  The distance metric can treat reliability as a weighting to reduce the relative contribution of *less reliable features*. In practice, it is difficult to adjust relative feature reliabilities, and it is not attempted here.

The remaining sections in this chapter describe each feature with its associated reliability measure.

Uniformly Reliable Cues

Some cues are uniformly reliable, e.g. the onset or offset of energy. Ideally, each such cue must be assigned a reliability such that the cue does not dominate other cues providing independent information. We tried assigning a value equal to the net mean reliability of the remaining cues.

However, experimental results did not show any benefit and this idea was discarded in favor of using all–ones reliability maps for such features.

Feature	Normalization	Reliability	Reliability normalization
Pitch	Standard normalization over $t\nu$ map	Harmonicity	5-95%ile range scaled to [0.0, 1.0]
ITD	Standard normalization over $t\nu$ map	Coherence	5-95%ile range scaled to [0.0, 1.0]
ILD	Standard normalization per band	All ones	–
Log Energy	Standard normalization per band	All ones	–
Common onsets	–	All ones	
Common offsets	–	All ones	
Log frequency	Standard normalization over $t\nu$ map	All ones	

Table 4.2: Feature and reliability computation scheme

4.2 Energy Map

Energy at a $t\nu$ pixel is computed as the logarithm of the energy localized in a narrow time window in that frequency band.

$$\mathbf{En}(c,t,\nu) = -10\log(\sum_{\delta=-T}^{T} w(\delta)|s(c,\nu,t+\delta)|^2) + Adj(\nu)$$

The frequency bands ν are centered at logarithmic values, and the signal decomposition into frequency bands by the gammatone filterbank is described in section 3.3.

A cosine squared time window is used for this computation to reduce high–frequency artifacts from pixel to pixel along time.

A random signal contains energy equally distributed in all frequencies, and can therefore be used to determine an additive adjustment factor for per band energy normalization. This makes the energy levels comparable across bands.

This approach ignores the instantaneous spectral profile of the signal, and assumes that the magnitude of signal energy is the same across all frequency bands dominated by that signal.

The log energy map is normalized to have zero mean and unit variance across all $t\nu$ pixels.

An all–ones mask is used as reliability for the energy map.

4.3 Energy differentials

Instants in time when the energy content in the frequency band is greater than the recent moving average by more than 50% are marked as pixels energy onsets. To allow grouping with pixels having onset within a common short time window, the onset value decays exponentially along time.

Similar to onsets, the offsets in energy are also tracked computed by thresholding sudden fall in energy, and

tracked by exponential decays.

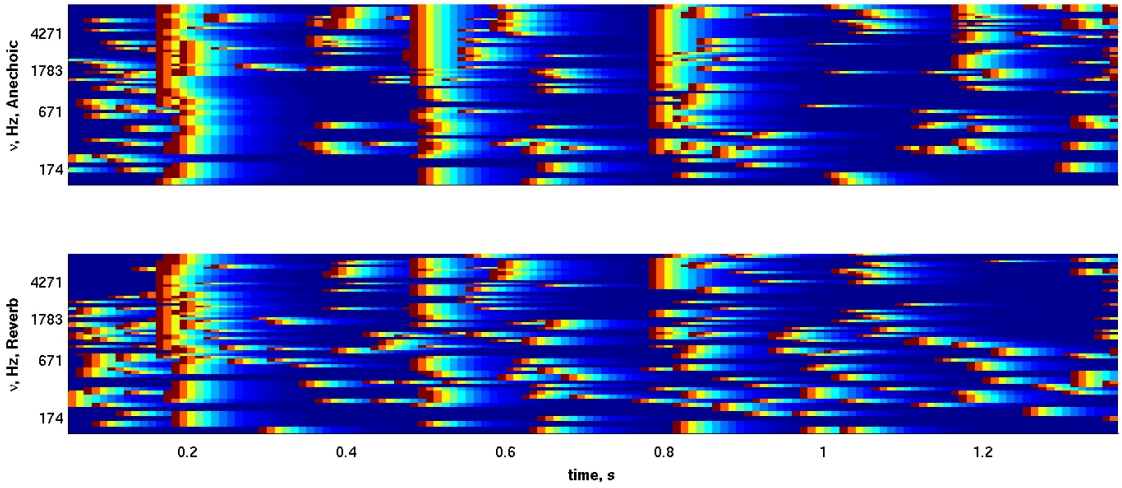


Figure 4.2: Energy onsets in an anechoic room (top) and for the same source mixture in a reverberant room (bottom). Note the fewer onsets in reverberation.

4.4 Frequency Band

The center frequency of each frequency band also serves as a feature, since neighboring frequency bands are more likely to belong together. This is facilitated by creating a uniform frequency map along time as follows –

$$\mathbb{F}_{t\nu} = \log \nu, \forall t$$

Since the center frequencies are logarithmically spaced, this feature assumes linear values along the frequency axis. This feature is also assumed to be uniformly reliable.

4.5 Pitch

Pitch is an ensemble characteristic defined per individual harmonic sound source, instead of being defined for a signal. A signal with multiple active sound sources therefore has multiple pitches, each corresponding to the respective source.

For instance, when two speakers, Alice and Bob, utter something at the same time, it makes sense to define

the pitch of Alice' and Bob's voices, but not for the mixed signal. Thus, knowledge of separated individual speakers in a mixture is seemingly a pre-requisite for computing their pitch.

4.5.1 Computational Models

Computationally, two distinct indicators of the generating vibration are available:

- The per band modulation of signal at the frequency of the generating vibration. This is exploited by *temporal algorithms*, that – in very general terms – infer modulation frequency by picking the autocorrelation peak. Temporal refers to the significance these algorithms place on the phase of the received signal, as opposed to ignoring it altogether as done by the spatial algorithms described next.
- The presence of energy at any given instant at a fundamental frequency and its harmonics. Given the spectral profile, the periodicity in the profile can indicate the pitch of the signal, which forms the basis of spatial models, so named because of their observing the location along tonotopic axis.

Both these indicators are utilized by *spatio-temporal algorithms*, that perform per-band auto-correlation, and combine information across bands to determine the final pitch. The algorithm used for computing pitch here belongs to this class.

In time instants dominated by unvoiced speech or silence, the computed quantity does not correspond perceptually to a *pitch*. Regardless, for ease of reference, this feature is referred to as pitch (denoted by Π) here.

4.5.2 Algorithm for $t\nu$ pitch assignment

A popular approach in literature is to attribute to a source all the harmonics of its pitch. To create a $t\nu$ map in this manner, a candidate pitch value can be computed at each instant in time, and all $t\nu$ pixels of all harmonics are assigned that value. For multiple pitches, this approach is modified in two subtle ways, similar to Assmann and Summerfield (1990), with steps listed in algorithm 5:

- Multiple candidate pitches are computed at an instant from the summary autocorrelation function. Computationally, these are the delays at which the top few percentiles of the values assumed by the summary autocorrelation occur.
- Each frequency band is assigned a value from amongst the candidate pitch values, determined from the autocorrelation function at that band. This value corresponds to the delay at which the ACF achieves the maxima from amongst the candidate pitches.
  - Considering only the candidate pitches is expected to retain peaks from actual vowels and eliminate the periodic peaks in the ACF.

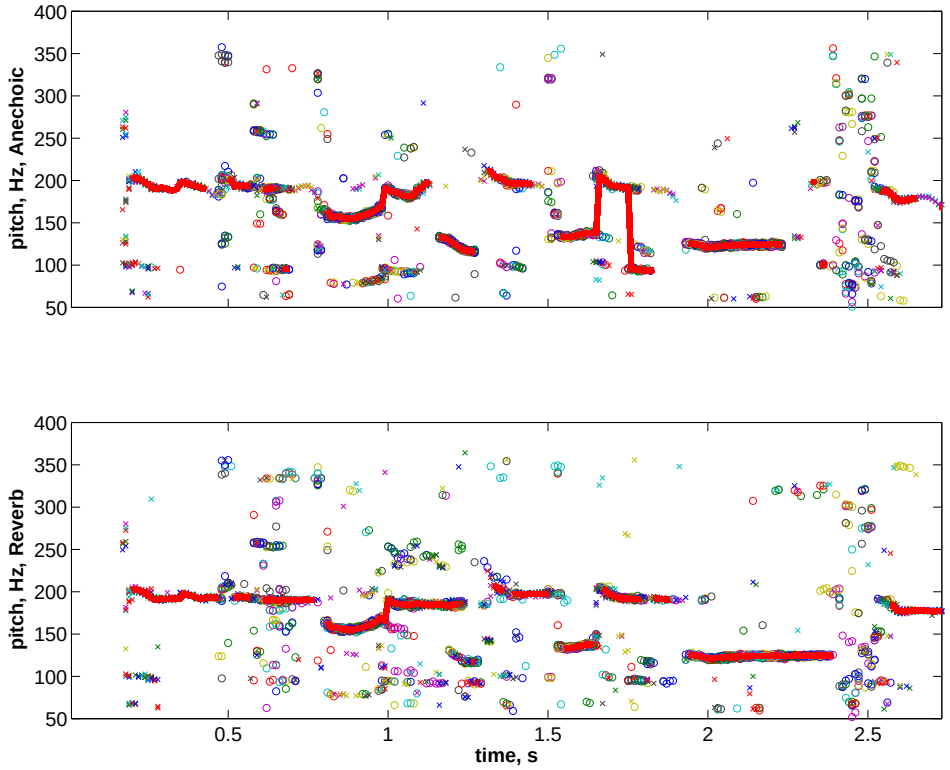


Figure 4.3: The per–band pitch computation algorithm (dots) compare to per–instant pitch computed from `praat` (overlaid continuous lines in red) in an anechoic chamber (top), and for the same source in a reverberant room. Mutiple pitch tracks are visible in the per–band pitch values while the pitch computed by `praat` flips between them.

Parameter		Value
Range of plausible pitch	τ	40 – 500 Hz
Autocorrelation width	δ	$\max(3\nu, 10\text{ms})$
Smoothing window for pitch	$w(t)$	cosine squared
Summary autocorrelation threshold for candidate pitches	k	90%ile

Table 4.3: Parameters for pitch algorithm

– Considering all candidate peaks prevents assignment of all bands to the dominant pitch, allowing appropriate source attribution when multiple speakers are simultaneously active.

Since only the dominant pitch gets picked at any $t\nu$ pixel, this procedure relies on a source dominating at least some of the pixels to enable separation.

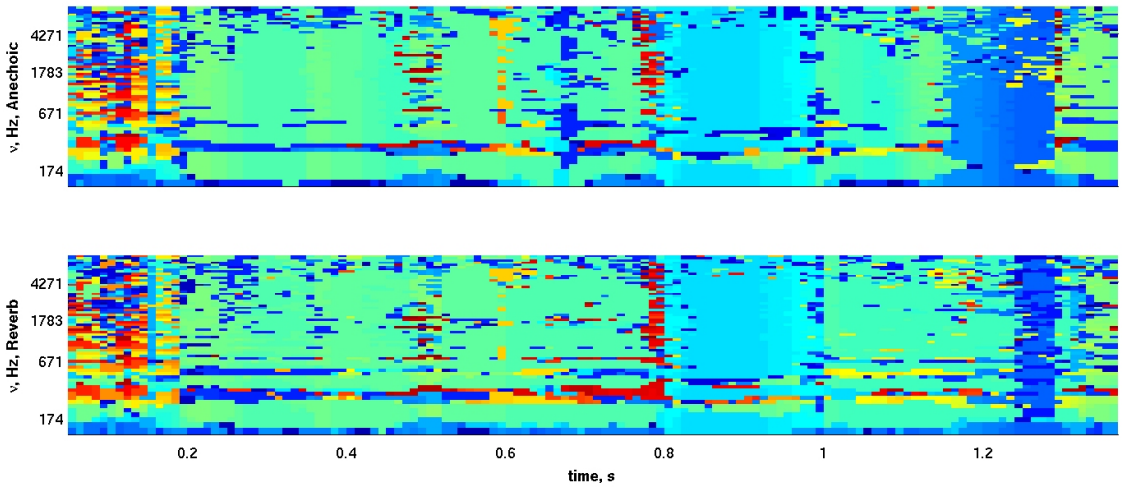


Figure 4.4: Pitch map computed in an anechoic chamber (top) and a reverberant classroom. Note the increase in variability in reverberation, and the smearing of pitch along time.

The information in the autocorrelation function at each $t\nu$ pixel summarizes influence of multiple pitches on that pixel. The information besides the dominant pitch that is discarded in the current approach can potentially be utilized for better separation. However, this idea is not developed further.

The performance of the algorithm is compared to the pitches extracted from `praat` in figure 4.3, and the benefit of linear performance is demonstrated later in figure 5.11.

4.5.3 Harmonicity

The strength of the pitch in a band is again an ensemble property, and is computed as the peak of the summary autocorrelation at the picked offset. Harmonicity values are clipped beyond 5–95%ile values, and normalized to lie in $[0, 1.0]$.

4.6 Interaural Time Difference (ITD)

The narrow band signals detected at the cochlea provide two distinct methods of decoding interaural time difference:

- The interaural phase difference (IPD) between the signal received at the left and the right ear. Within a band, this has the disadvantage of being unique only upto a phase shift of 2π, or roughly $\frac{2\pi}{f_c}$ seconds, where f_c is the center frequency of the band.
- By combining information across multiple frequency bands, it is possible to determine the common

Parameter		Value
Range of plausible ITD	τ	[-1ms, 1ms]
Cross–correlation width	δ	$\max(3\nu, 10\text{ms})$
Smoothing window for ITD	$w(t)$	cosine squared

Table 4.4: Parameters for ITD algorithm

interaural time difference (ITD) that can explain all the observed IPDs. Thus, ITD is an ensemble feature of $t\nu$ pixels at an instant, and is amenable to computational treatment quite similar to the computation of pitch.

ITD is comparable across time and frequency bands, and is therefore used here as the measure of interaural signal delay.

Algorithm 6 describes the ITD selection algorithm.

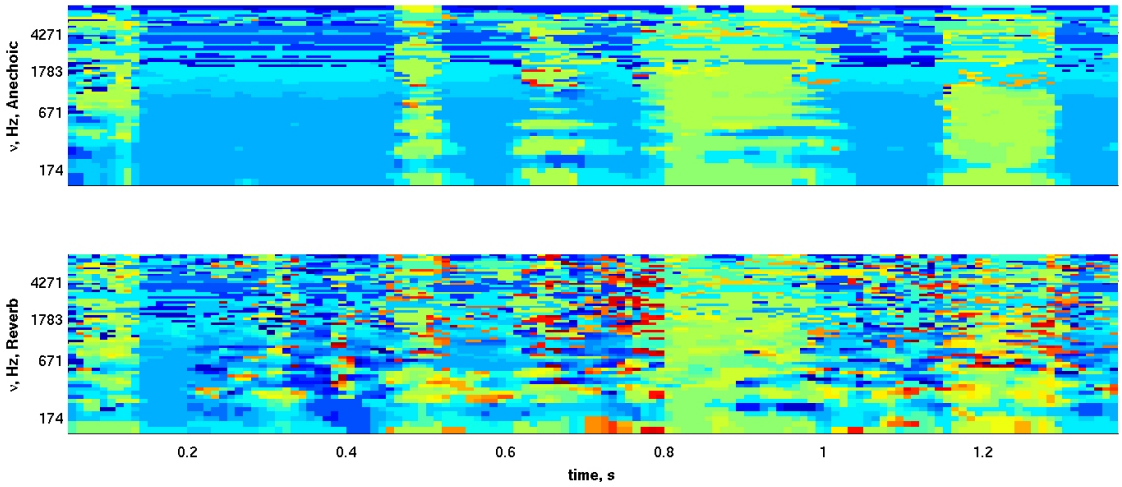


Figure 4.5: ITD map computed in an anechoic chamber (top) and a reverberant classroom for the same source. Note the increase in variability in reverberation.

The peaks of the summary cross–correlation function correspond to the ITDs of the dominant sources. The position of the receivers, determines the range of interaural delay that can physically exist, and for the algorithm, assuming that the signals are picked up by sensors located at human ears, this range is set to $[-1ms, 1ms]$. In this range, it is unlikely that multiple peaks for the same ITD exist for sources as wide–band as human voice. Therefore, each peak of the summary cross–correlation are more likely to be at delays corresponding to ITDs of the active sound sources.

In a frequency band centered at f_c, the cross–correlation function has multiple peaks, corresponding to

- periodic peaks at delay intervals of $\frac{1}{f_c}$, and
- strong peaks corresponding to interaural delays of the sources.

Emphasizing the cross–correlation function by multiplying each value with the summary cross–correlation subdues the periodic peaks leaving behind only the peaks corresponding to interaural delays of the source. The peak of the emphasized cross–correlation is more likely to be the one corresponding to the stronger source at the instant. However, we need to pick the ITD corresponding to the source dominating the individual band, and therefore, consider the peaks of emphasized correlation as the candidate peaks.

To obtain the ITD of the source dominating the band, the per-band cross-correlation is masked with the candidate peaks, and the location of peak is extracted.

4.6.1 Binaural Coherence

Similar to harmonicity for pitch, binaural coherence, which is the height of the per band interaural cross–correlation at the chosen ITD delay, is used as the reliability for ITD.Coherence values are clipped beyond 5th and 95th percentile values, and normalized to lie in $[0.0, 1.0]$. Figure 5.5 shows using coherence as weighting over ITD allows nearly as good separation of sources as the best linear separator (for frequency bands upto 2 KHz).

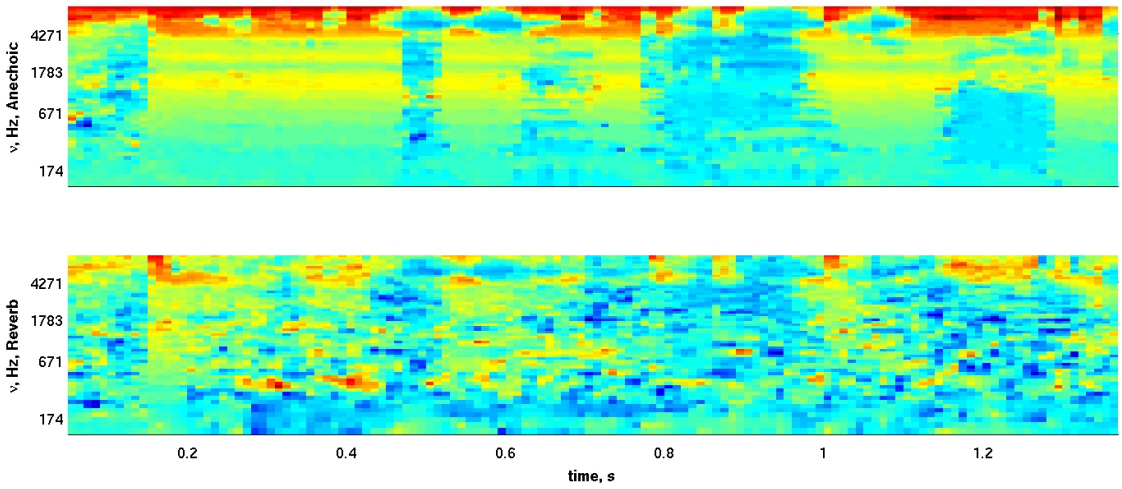


Figure 4.6: ILD maps computed in an anechoic room (top) and a reverberant classroom (bottom).

4.7 Interaural Level Difference (ILD)

The difference in energy maps of the left and the right channels defines the interaural intensity difference. Since energy is the logarithm of instantaneous RMS value of the signal around the time instant, this quantity

is the logarithm of the ratio of the RMS values from the two channels.

$$\mathbb{L}^{L/R}(\nu, t) = 10 \log_{10} \mathbf{En}_{L,\nu,t} - 10 \log_{10} \mathbf{En}_{R,\nu,t}$$

where $\mathbf{En}_{(.,\nu,t)}$ is the energy energy content at time t, frequency band ν, and channel L or R.

Per-band ILD varies with frequency and energy content. To make ILD values comparable across all frequency bands, a per-band offset is applied. This offset is computed by averaging for a given band the ILD for white noise at multiple locations in the room, generated using HRTFs.

An all–ones map is used as the reliability for ILD.

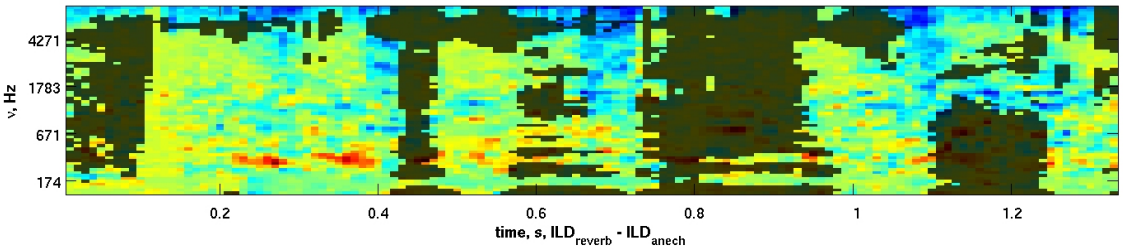


Figure 4.7: Difference in ILD maps in anechoic and reverberant rooms shown in figure 4.6 highlighted with the region dominated by a single source.

Chapter 5

Developing The Segmentation-Grouping Algorithm on a two-speaker mixture

5.1 Evaluation setup during development

Stimuli

The Segmentation-Grouping Algorithm is evaluated on simulated speech mixtures. Everyday sentences in American English spoken by eight male talkers eight female talkers were selected at random from the TIMIT acoustic-phonetic corpus. The sentences were normalized to have the same energy relative to each other, and then convolved with HRIRs to spatialize them. HRIRs measured with a KEMAR dummy head in the middle of a classroom (T60$=$ 0.87s) and a chapel (T60$=$ 1.07s) are used to simulate reverberant mixtures. For simulating anechoic conditions, the classroom HRIRs are truncated to eliminate reverberant energy, with a resultant T60 of 0.02s.

Iterative versions of speech separation algorithms are evaluated in two talker (*female*/male, target *emphasized*) conditions, with target set right ahead of the listener, and the distractor to the right separated by angles from 0° to 90° in increments of 15°. This makes left ear the *better ear* in all conditions where sources are spatially separated; and in the results here, the best algorithm would be the one that maximizes enhancement at the left ear.

Twelve mixtures are constructed randomly to evaluate each condition.

Evaluation metric

The enhancement in sound to distortion ratio (SDR) in going from the mixture spectrogram to the masked spectrogram is used as the primary metric for comparison. This metric is described in Vincent et al. (2005a), adapted for computation from $t\nu$ masks as follows:

$$SDR_{enh} \quad = \quad 10\log_{10}\frac{\mathcal{E}_{\tilde{s}}}{|\mathcal{E}_{\tilde{s}}-\mathcal{E}_{\hat{s}}|} - 10\log_{10}\frac{\mathcal{E}_{\tilde{s}}}{|\mathcal{E}_{\tilde{s}}-\mathcal{E}_{m}|}$$

Here, $\mathcal{E}_{\tilde{s}}$ is the energy of the source reconstructed from the veridical fractional mask, $\mathcal{E}_m$ is the energy in the mixture, $\mathcal{E}_{\hat{s}}$ is the energy in the reconstructed source. With this metric, an all–ones mask shows 0 dB SDR enhancement in all conditions, the veridical binary mask (VBM) shows the limit of performance of all binary reconstruction masks. The perfect mask, which is the ratio of target to mixture energy, produces ∞ dB SDR enhancement.

5.2 Baselines

Figure 5.1 shows the baselines of separation performance against which a speech separation algorithm may be evaluated.

5.2.1 Ideal Performance

Ideal $t\nu$ Pixelwise Binary Assignment

The ideal target to masker energy ratio mask yields infinite gain in SDR. Rounding this ideal fractional mask yields the ideal binary mask. Separation using the ideal binary mask is the upper bound on performance of binary masks that attribute the energy content of a $t\nu$ pixel exclusively to either source. This upper bound denoted in the figures by the legend VBM is the SDR of the reconstructed target signal $\hat{T}$ from the mixture Msuch that:

$$\hat{T}_{\text{ibm},\nu,t} = k_{\text{ibm},\nu,t}.M_{\nu,t}$$

where $k_{\text{ibm},\nu,t} = \left\lfloor \frac{\mathbf{En}_{\nu,t}}{\mathbf{En}_{in,\nu,t}} + 0.5 \right\rfloor$.

Note that $\lfloor . + 0.5 \rfloor$ is the rounding function, since the ratio of energy content is a positive quantity.

Discussion

- *Effect of space* For the better ear, $\Delta SDR$is best for 0° azimuth, since the original SDR is high enough as the azimuthal separation increases, leaving lesser potential for improvement. The opposite effect of increasing azimuthal separation is observed in the worse ear, and ΔSDRincreases marginally with angle.
- *Effect of reverberation* Performance falls as reverberation increases while moving from anechoic, to classroom, to Marsh chapel situations. For the worse ear, the fall in performance is marked due to much lower SDR in the original mixture; whereas performance decline is not as dramatic in the better

ear, where the target source is dominant.

5.2.2 Worst Case Performance

No Separation

The no separation case, where the output signal is the same as the input signal, has no improvement in SDR, and is not demonstrated.

Equalization of Mixture Energy to the Target

The SDR measure is aware of the energy levels of the reconstructed source, hence, it is possible to get a change in SDR only by rescaling the input mixture. It is analytically difficult to determine for the scaling that maximizes the ΔSDR with constant scaling applied to the input. The root mean square equalization between the input mixture and the ideal target gives a constant scaling that would minimize the mean square error between them:

$$\hat{S}_{\text{EnEq},\nu,t} = k.S_{\text{in},\nu,t}$$

where $k = \dfrac{\text{RMS}_\text{S}}{\text{RMS}_\text{in}}$.

This sanity check denoted by **EnEq**, is the lower baseline representing the performance of a no-op algorithm when it is greater than 0 dB.

Discussion

- *Effect of space* For the better ear, at 0° separation, scaling down the mixture energy raises the ΔSDR metric value. But as the angular separation increases, the target source becomes dominant, and scaling it down removes more of target energy than the masker. For the worse ear, this effect is reversed.

- *Effect of reverberation* Reverberation smears the energy of target and masker across the spectrogram, causing increased mixing of the sources. Scaling down the mixture energy therefore removes more energy attributable to the masker, than to the target. The net effect is therefore generally non-zero improvement in SDR. The no-op baseline in cases with negative ΔSDR is 0 dB.

The energy equalization is the best amongst the worst case separation, and should be treated as the no-op baseline for separation performance when positive.

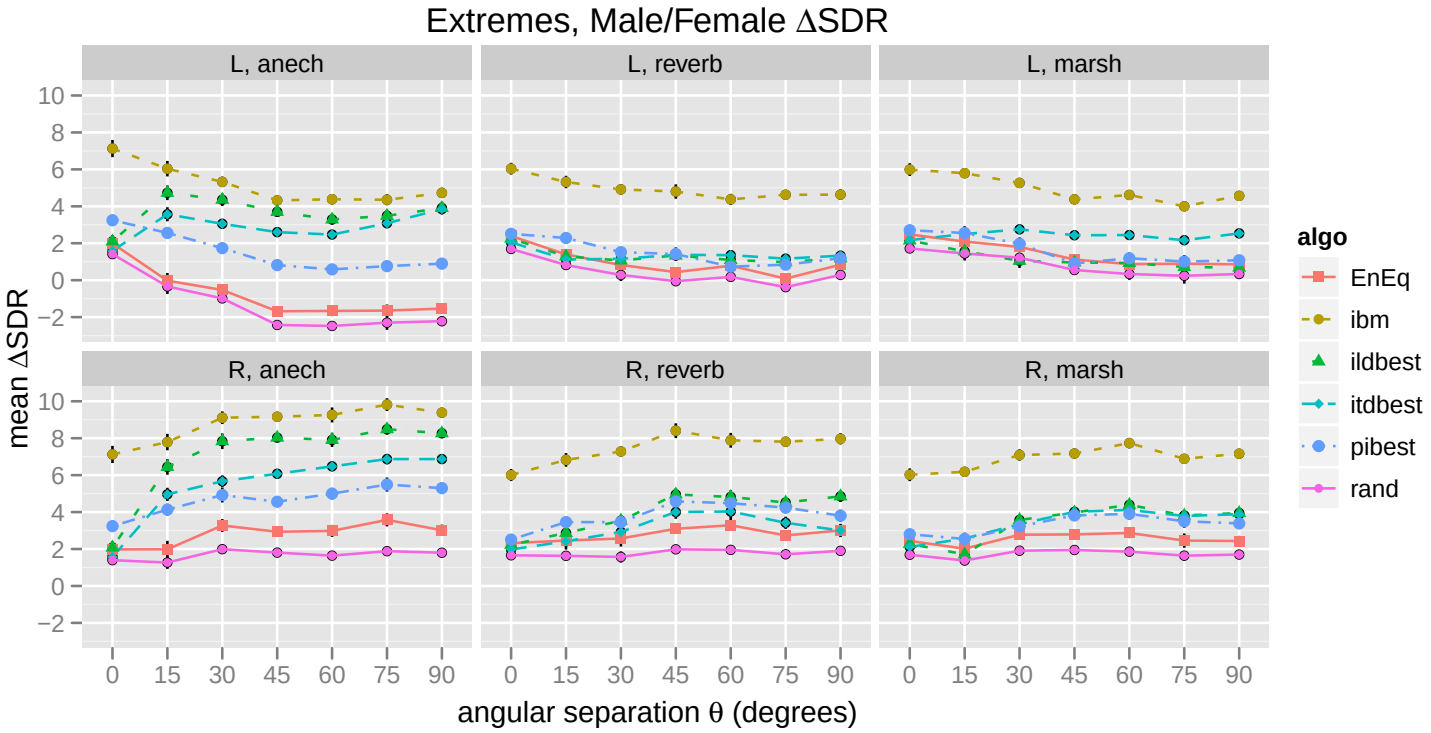


Figure 5.1: Baselines for all figures. *Female*@0°/Male@α° at 0 dB SPL at the source, with the female voice right in front, and the male voice separated by azimuthal angle indicated along the x axis. For anechoic case, T_{60} is $0.02s$; the reverberant case is a classroom with early echoes and T_{60} of $0.87s$; and the highly reverberant case is from Marsh Chapel with T_{60} of $1.07s$. Lines are described in table 5.1

Random Masks

The speech separation problem is particularly challenging when the SDR is low in the original mixture. For such mixtures, a sanity check can be performed by comparing the performance of any algorithm against a random mask. The random mask considered here assumes equal energetic contribution from each source, and assigns each $t\nu$ pixel randomly to the target or masker with equal probability. Multiple such masks are generated for a given input mixture, and ΔSDR is averaged and reported as ΔSDR_{rand}.

$$\hat{S}_{\text{rand},\nu,t} = p.S_{\text{in},\nu,t}$$

where

$$p \sim \text{Bernoulli}(\tfrac{1}{2}).$$

Discussion The energy equalized mask is better than the random mask since the former creates a mask that is correlated with the target at least as much as the origianl mixture. The latter eliminates energy randomly from the mixture, and makes decorrelates the reconstructed signal further from the target than the original mixture.

5.2.3 Ideal Linear Separation by Individual Features

Linear separation of sources occurs when a $t\nu$ pixel is labeled as target or masker based on its position in feature space relative to a linear separating surface.

For a single feature, the separating surface is a threshold value, and all $t\nu$ pixels with feature value above the threshold are assigned to the target, and the rest are attributed to the masker. To determine the threshold that maximizes ΔSDR, a linear search is performed by trying out all values as candidate thresholds. The resulting best ΔSDR is presented in the figures for ITD, $\mathbb{L}^{L/R}$, and Π features as **bestitd**, **bestild**, and **bestpi** algorithms respectively.

Discussion

- *Effect of space* The linear separation by ITD and ILD is significant as the azimuthal separation becomes non-zero. For the better ear, the improvement plateaus quickly beyond 15°, because the target is dominant in the initial mixture and the initial SDR is high enough to leave little scope for improvement. For the worse ear, the improvement with increasing angular separation is more pronounced.

  Pitch helps in separation at 0°, and does not benefit from spatial separation as expected. The trend in **bestpi** performance with angle demonstrates the effect of SDRin the mixture on achievable separation performance. Specifically, for the better ear, initial SDR increases with angle leading to reduced scope for improvement; and the effect is reversed for the worse ear.

- *Effect of reverberation* Reverberation is unforgiving toward ILD. The change is noticeable in the histogram of source ILDs in figure 5.2, as well as in the ILD maps from figure 4.7. In the anechoic case, ILD is is the best cue for separation, but moving to a room with $T_{60} = 0.87s$ obliterates the ability to use ILD as a separation cue in the better ear. Moving to the Marsh Chapel, a reverberant room with longer T_{60} $(1.07s)$ further worsens the near-random performance of **bestild**. In the worse ear in reverberation, **bestild** performs slightly better than random, presumably because of the better scope available for separation due to lower initial SDR.

  Performance of **bestitd** also degrades in reverberation. Interestingly, the performance is worse for the classroom that has shorter T_{60} , than the chapel. This may be due to early echoes in the classroom, as shown in figure 2.1 that compares the reverberant activity in the two rooms.Such reflections have ITD of the reflecting surface, making them prone to misclassification, and significant chunks of energy, making them impactful on the SDR metric.

  The performance of **bestpi** does not change much, but the changed baseline in reverberation eats into the gains due to pitch.

  Thus, linear separation by individual features is of limited use in reverberant situations.

  Different types of reverberation is present in the two rooms, and the fall in performance is not predicted by T_{60} alone 2.1.

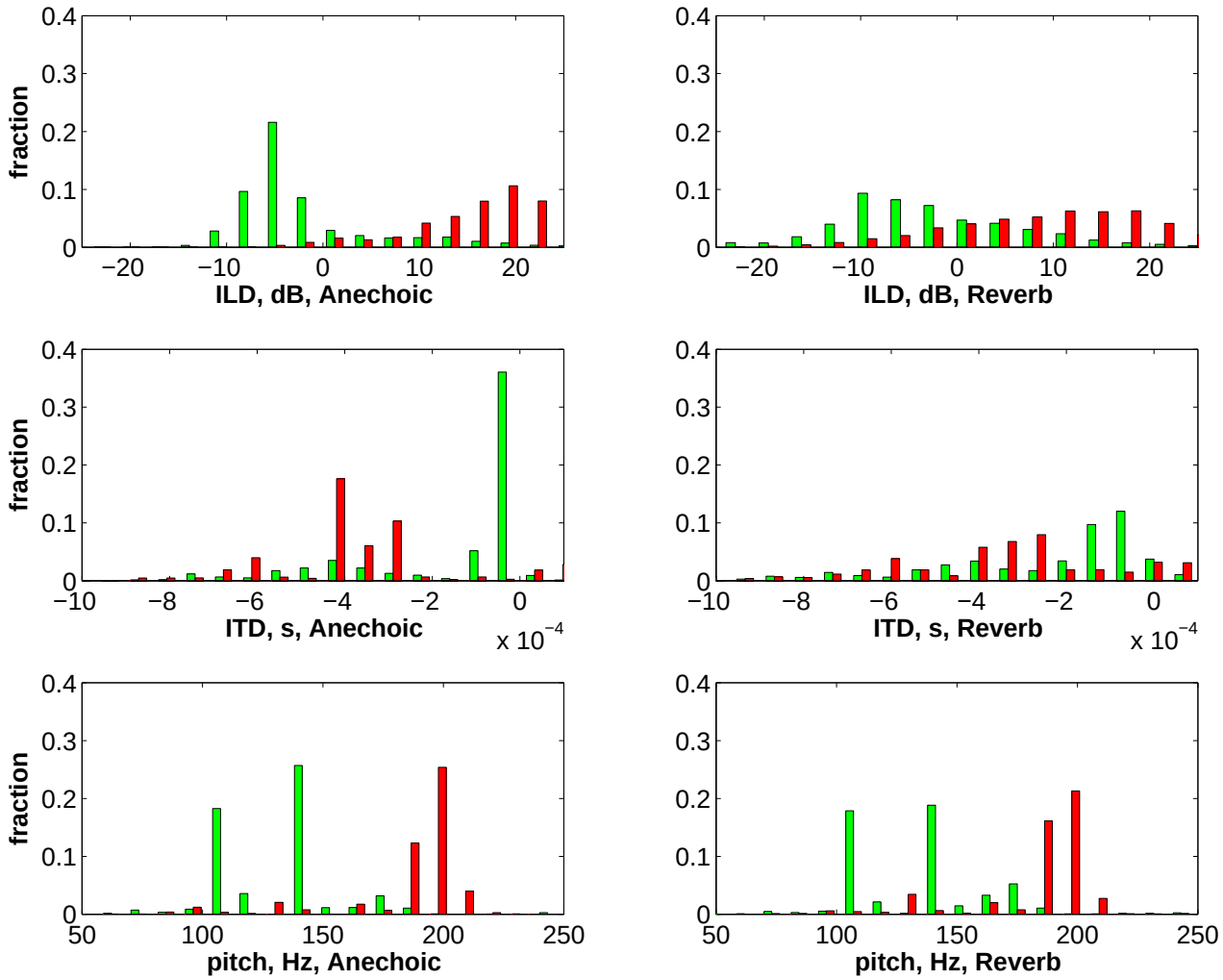


Figure 5.2: Effect of reverberation on feature histograms. Red and green bars represent fractions of energy belonging to either source in a two talker mixture. The anechoic room has T_{60} = 0.02 s, and the reverberant room is the classroom with T_{60} = 0.87 s.

Type	Mnemonic	Description
Baselines	VBM	Ideal binary mask that scales $t\nu$pixels dominated by the target to 1.0, and the remaining pixels to 0.0. This is a veridical mask.
	EnEq	Energy Equalization mask that scales all pixels by a constant factor to equalize RMS of the mixture signal to the target signal RMS. This is a veridical mask.
	rand	Average performance of 20 masks, for each of which, each pixel is randomly assigned to target or masker with equal probability.
Linear Veridical	**bestild**	Best linear separation by ILD determined by sweeping over the range of ILD and picking value that yields maximum ΔSDR.
	bestitd	Best linear separation by ITD, determined by sweeping over the range of ITD and picking value that yields maximum ΔSDR.
	bestpi	Best linear separation by F_0, determined by sweeping over the range of pitch and picking value that yields maximum ΔSDR.
	pibestPraat	Best linear separation using the F_0 map computed from `praat`.
Linear Supervised	**ildglm**	Best logistic regression with ILD, trained and tested on the same mixture signal.
	itdglm, **itdglmCoh**, **itdglmEn**	Best logistic regression with ITD, trained and tested on the same mixture signal. The three versions have no weighting, coherence map, and energy map as weighting respectively.
	piglm, **piglmHar**	Best logistic regression with F_0, trained and tested on the same mixture signal. The versions are without any weighting, and with harmonicity as weighting respectively.
	piitdildglm	Best logistic regression with F_0, ITD, and ILD, trained and tested on the same mixture signal.
Segmentation Baselines	**seg0Lin**	Ideal linear separation by ITD, of segments created without using feature reliabilities.
	segHCLin	Ideal linear separation by ITD, of segments created using harmonicity and coherence as feature reliabilities for pitch and ITD respectively.
	segHCIBM	Ideal binary separation of segments that incorporate feature reliabilities. The difference of this algorithm from VBM represents the cost of treating pixels as aggregate segments.
Segmentation Final	**steppeSpace**	Segmentation–Grouping algorithm output when run with only spatial cues.
	steppeSpatial	Segmentation–Grouping algorithm using all cues during segmentation, but only spatial cues during grouping.
	twotierHCEn, **steppeSegHarCohEn**	Segmentation–Grouping algorithm output when run with all cues mentioned in table 4.2. The two are variants of grouping algorithm.
	steppeUr	Monolithic clustering over all pixels into two groups, one of which is picked as the source. The gain of **twotierHCEn** over this represents the benefit of adding the complexity of sequential segmentation before grouping.
	steppeSeg0	Clustering with no reliabilities in the segmentation stage.

Table 5.1: Description of algorithm mnemonics used in figure legends in this chapter.

5.3 Baseline Performance in lower frequency bands

Of the 64 spectral bands, it is instructive to see the breakdown of ΔSDR performance in the lower 48 (upto 2 KHz), and the higher 16 spectral bands (2 - 8 kHz) separately. Correlation based cues (pitch and ITD) are more robust in the lower bands compared to the higher frequency bands, and are expected to yield better performance below 2 KHz.

In contrast, ILD cues are available for higher frequency bands, since the head blocks high frequency components more effectively but is transparent to lower frequency components.

Figure 5.3 shows the effectiveness of linear separation by ITD and pitch in lower frequency bands.

Increase in reverberation (on moving from anechoic chamber to classroom and Marsh chapel) erases all gains due to ILD in these frequency bands, presumably due to filling in of energy from all directions due to reflections. ITD gains are also markedly reduced; while linear separation by pitch remains the most resilient to reverberation.

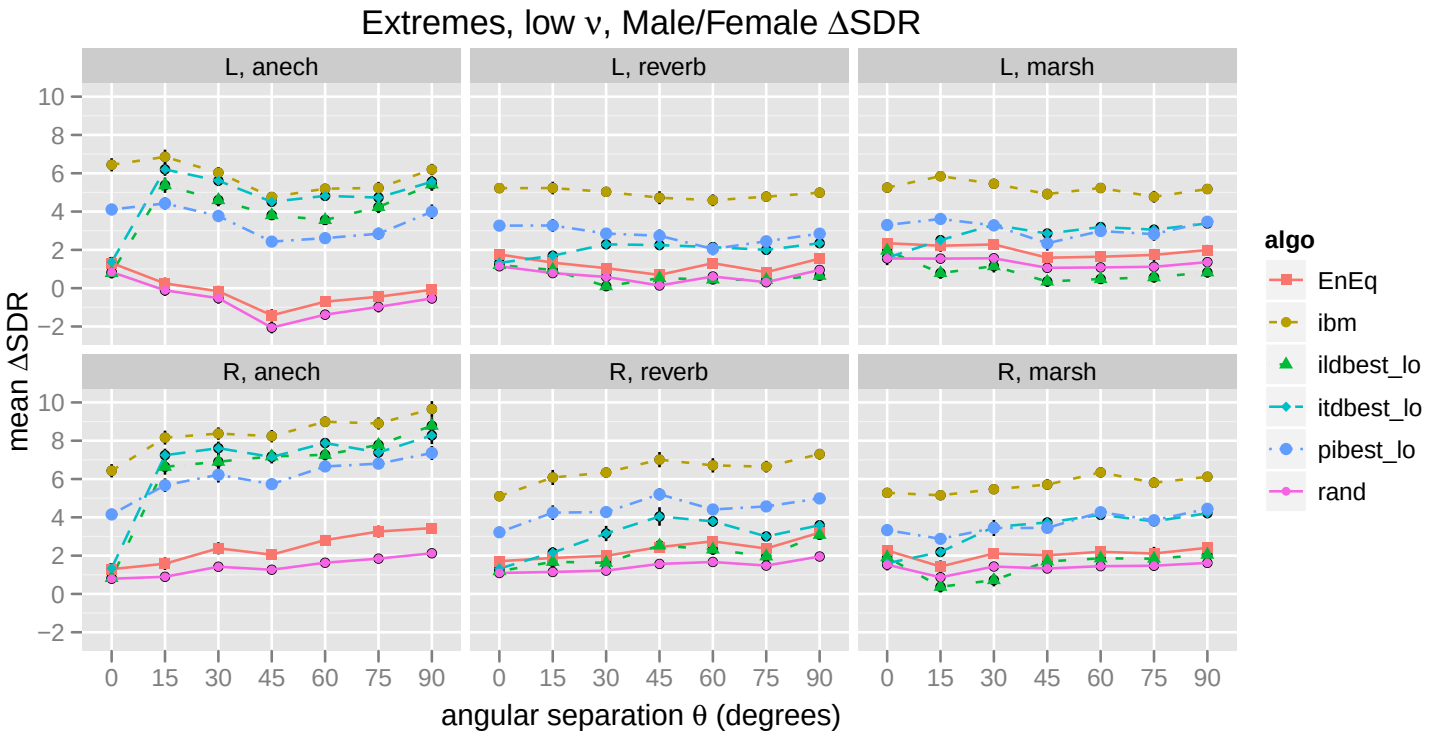


Figure 5.3: Baselines for low frequency bands

5.4 Baseline Performance in higher frequency bands

Figure 5.4 shows the effectiveness of linear separation by ILD in the higher frequency bands; and the limited utility of separation by pitch and ITD.

ILD at higher frequencies shows gradual decline with increase in reverberation, in contrast to the rapid decline in lower frequency bands.

Pitch and ITD also remain marginally usable, but as weaker cues than ILD, and quickly degrade in reverberation.

Discussion

- *Piecewise linear separation* Consider a single feature, say pitch, in a reverberant room. The best linear separation over the full spectrogram, shown in figure 5.1, is worse than the separate performance in figures 5.4 and 5.3. Hence, using separate thresholds for high and low frequencies, piecewise linear separation with a single feature can outperform linear separation with a single threshold.
- *Ensemble of features* Picking separate features (with associated optimal thresholds) for higher and lower frequency bands can potentially outperform using single feature with different thresholds.

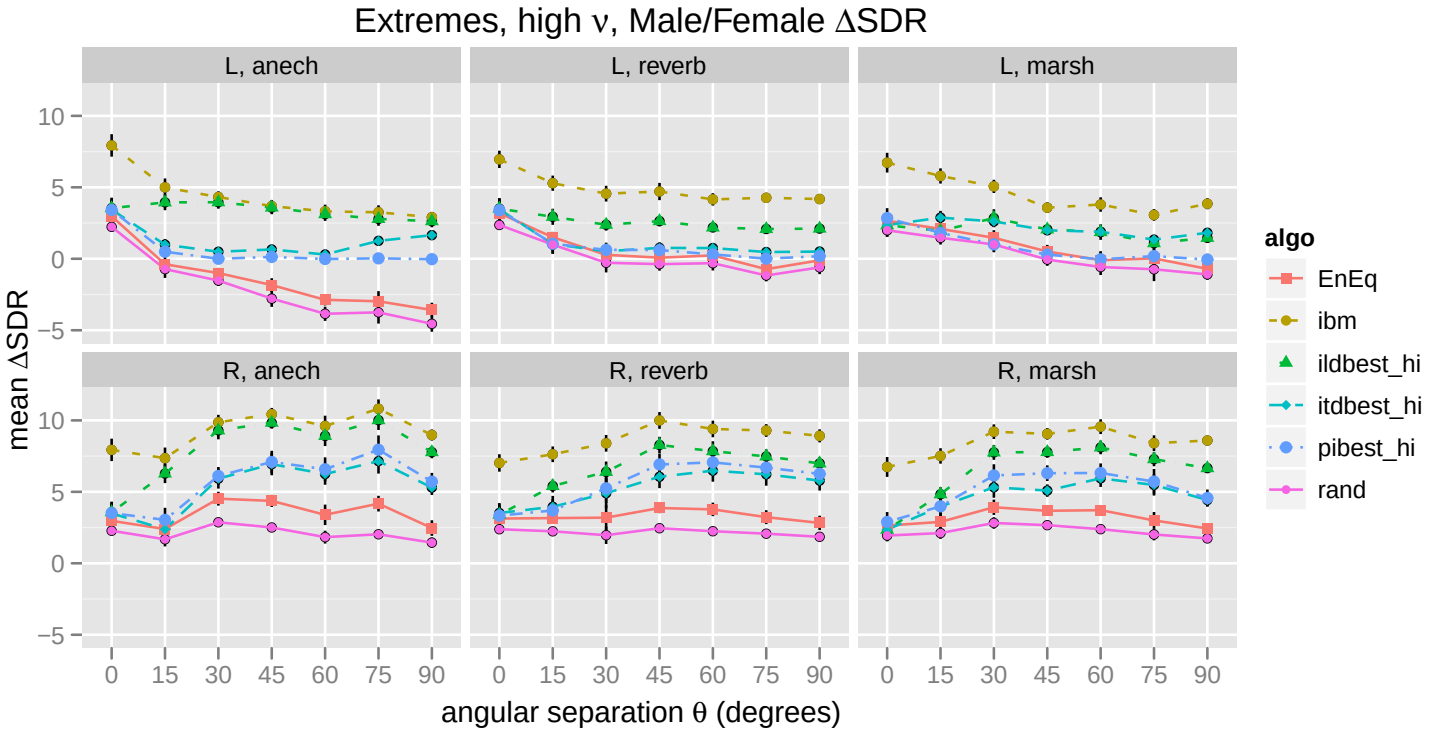


Figure 5.4: Baselines for high frequency bands

5.5 Separation by supervised learning with single features

The ideal linear separation discussed till is the upper bound for an unsupervised linear separator, since it makes no assumptions about the distribution of the feature values. To compare performance of our algorithm against a competing unsupervised algorithm, we train a linear discriminant, namely a logistic regression model implemented with Generalized Linear Models in MATLAB.

Feature value from the $t\nu$ map serves as the feature vector of unit dimension in case of single features discussed in this section. Training labels marking ideal classification of a pixel as target or masker is provided by the ideal binary mask. In most machine learning algorithms, a model trained thus is evaluated on a test dataset, since the performance on the training set itself is recklessly optimistic measure of generalization to new data. Here, the training set performance ΔSDR is presented in all figures to show the *bounds* on performance of supervised algorithms trained with the available features.

This algorithm is a proxy for performance of algorithms that perform linear separation in the space of a few features, such as ILD and ITD (Yilmaz and Rickard, 2004; Aarabi, 2002).

5.5.1 Linear Separation by ITD

Figure 5.5 demonstrates supervised learning with ITD as the single feature.

- **itdglm** The **itdglm**algorithm is a logistic regression model trained using ITD as the single feature, with each $t\nu$ pixel weighted equally. Performance is evaluated on the training labels themselves, and therefore the ΔSDR performance in figure 5.5 represents the upper bound on performance of a linear logistic model that uses ITD as the sole feature for separation. A logistic regression model assumes that the feature values have a normal distribution. Therefore, some performance is expected to be lost in comparison with **bestitd** since the latter is computed without any assumptions as to the distribution of ITD values.
- **itdglmCoh** Instead of weighting ITD at each $t\nu$ pixel equally, if the ITD values are weighted higher for pixels with higher binaural coherence in the logistic regression model, the performance improves. In figure 5.5 it is observed to be comparable to **bestitd** in most cases.
- **itdglmEn** To demonstrate that binaural coherence is higher not only for pixels with higher energy, a logistic regression model is trained with energy content (on logarithmic scale) per pixel as the weight. Figure 5.5 shows that while it improves performance compared to the unweighted baseline (**itdglm**), it is subpar to **itdglmCoh**.

Binaural coherence is demonstrated to be a good measure of ITD reliability, and henceforth used as the *reliability map* of ITDs. The application of reliability maps is described in later sections.

5.5.2 Linear Separation by ILD

Figure 5.6 demonstrates supervised learning with ILD as the single feature. Since ILD is available when either sound source is active in the mixture, ILD at all $t\nu$ pixels is weighted equally. The logistic model trained in this manner, denoted by **ildglm**, exhibits performance comparable to **bestild**.

5.5.3 Linear Separation by Common Harmonicity

Figure 5.7 demonstrates supervised learning with pitch as the single feature.

- **piglm**denotes a logistic model trained with pitch as the feature, weighting all $t\nu$ pixels equally.
- **piglmHar**denotes a logistic model trained with pitch as the feature, but with each $t\nu$ pixel weighted by the strength of harmonicity at that pixel. The performance in this case is seen to be close to **bestpi**, and hence, strength of harmonicity is used as the reliability map for pitch in subsequent work.

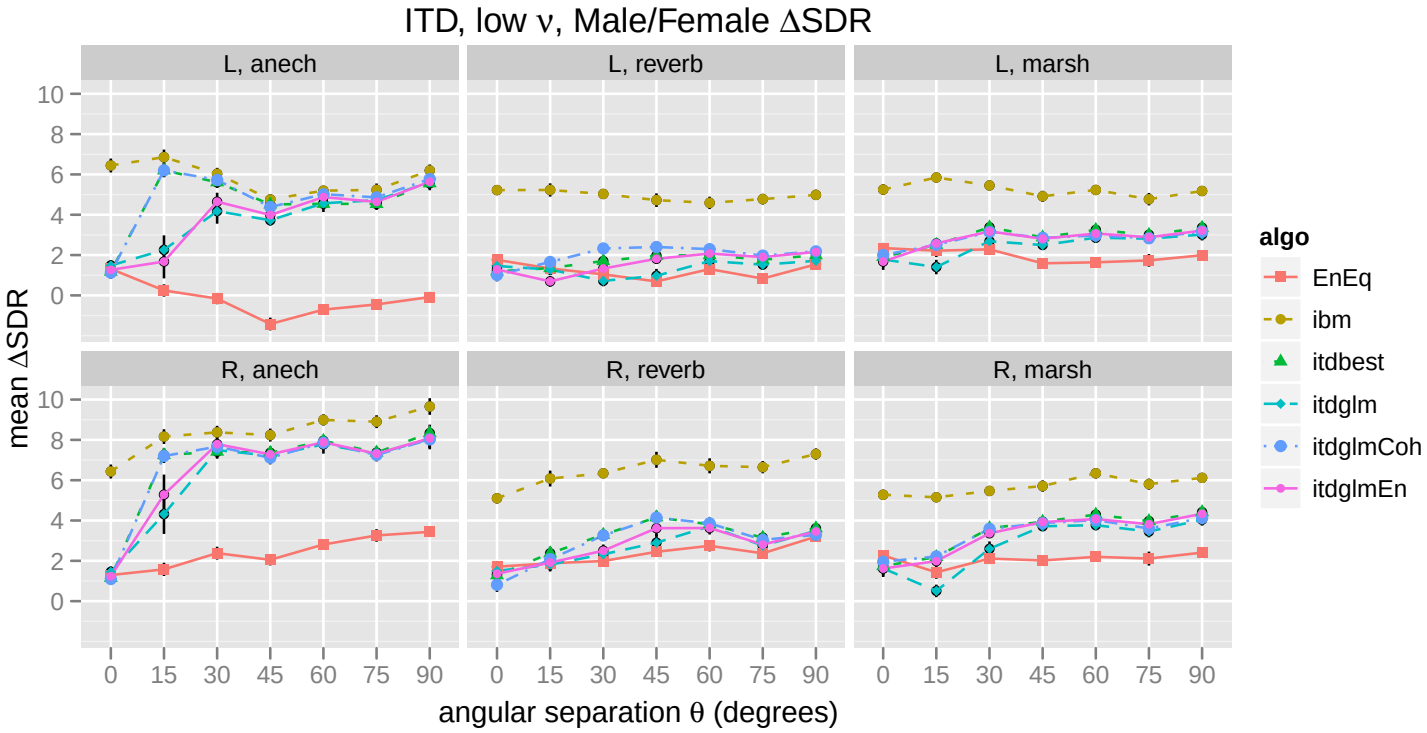


Figure 5.5: Linear separation by ITD, low

- *Evaluation in grouped frequency bands* In figure 5.7, the ΔSDR from lower frequency bands is shown. Note that the ΔSDR is *calculated* over the lower frequency bands, after the algorithms **bestpi** and **piglmHar** etc. have been run on the *full spectrogram*. Because of this, **piglmHar** outperforming the upper bound of **bestpi** in this figure is not an error; and happens because **bestpi** optimizes over the whole spectrogram, and is able to do so by performing a linear search over values of pitch. On the other hand, **piglmHar** is forced to optimize over lower frequency bands (since higher frequency bands have pitch values that are indiscriminable between the sources, especially, e.g. for fricatives).

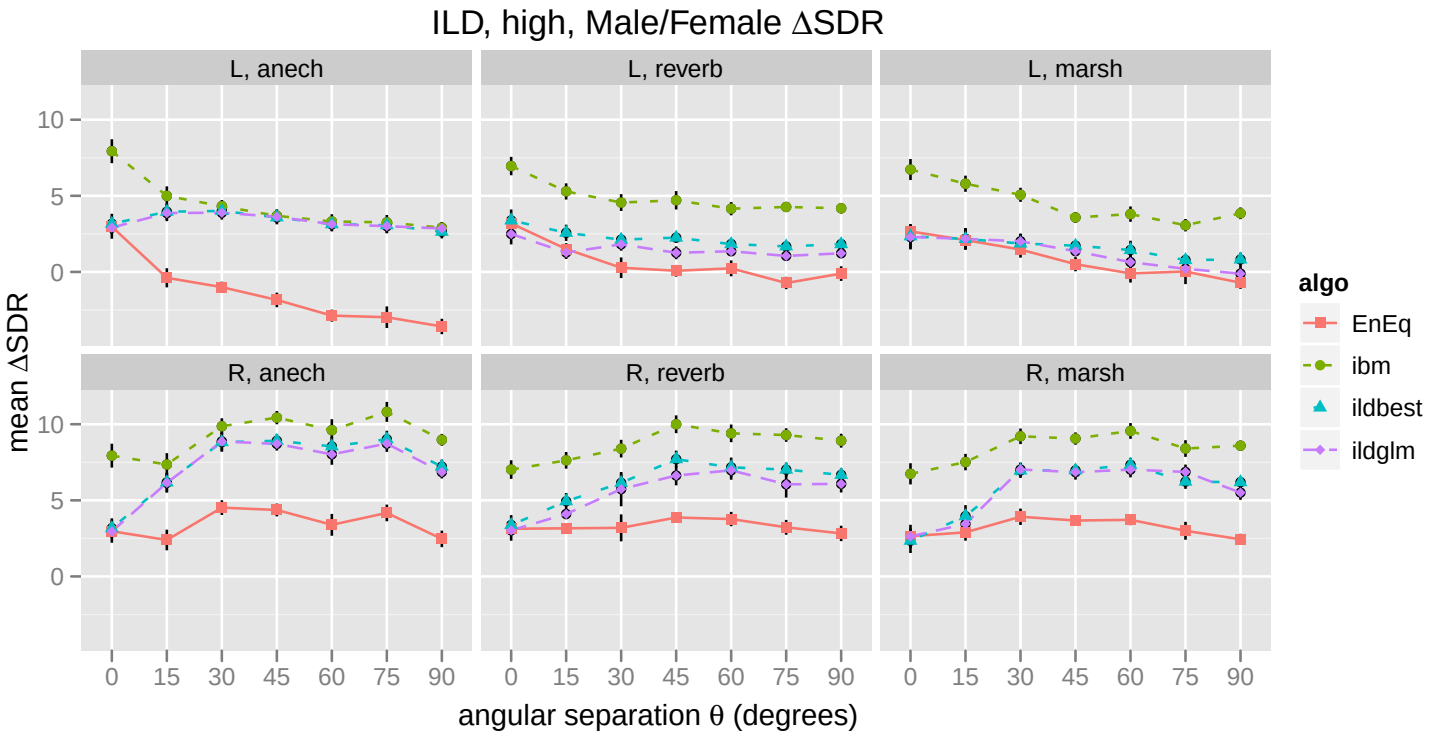


Figure 5.6: Linear separation by ILD, high

Figure 5.7: Linear separation by Harmonicity, low

5.6 Supervised Linear Separation by Multiple Features

In the previous section, logistic regression trained on individual features were demonstrated to perform close to best linear separation. In figure 5.8, a logistic regression model is trained on three features (ITD, ILD, and pitch), denoted by **piitdildglm**, and is shown to be at par with best linear separation by any individual feature alone.

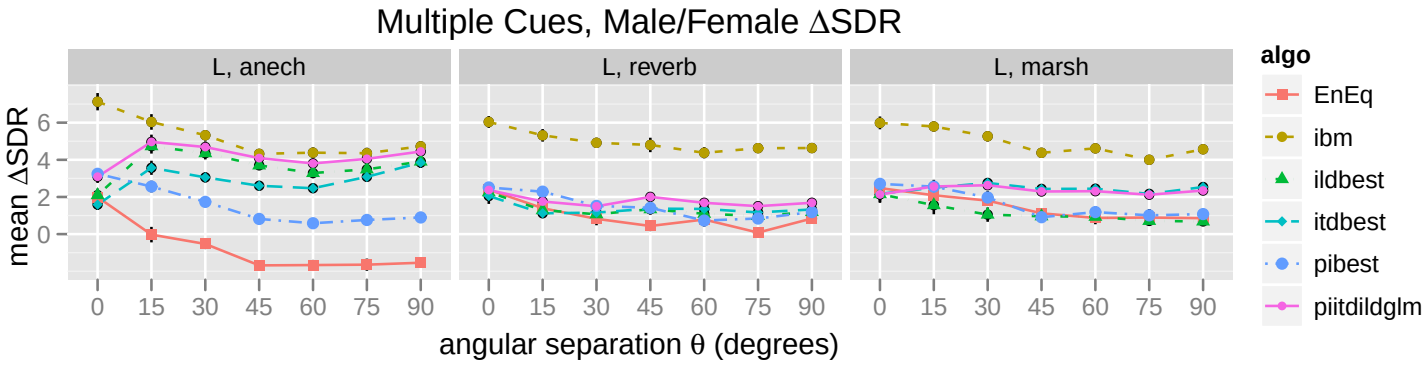


Figure 5.8: Linear supervised separation by multiple cues

Note that **piitdildglm** tries to optimize over the full spectrogram with all the features. In doing so, it underperforms versus **bestild**at high frequency bands (see figure 5.10, reverb, R), and underperforms versus **bestpi**in lower frequency bands (see figure 5.9, reverb, R). An ensemble algorithm tying separation by ILD at higher frequencies and ITD and pitch at lower frequency bands can potentially outperform **piitdildglm** in the overall spectrogram.

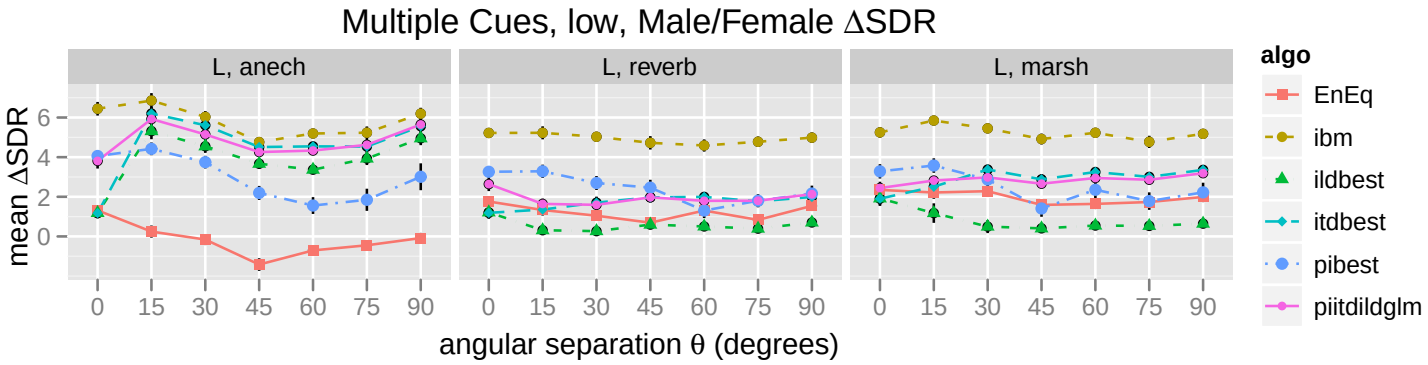


Figure 5.9: Linear supervised separation by multiple cues, low

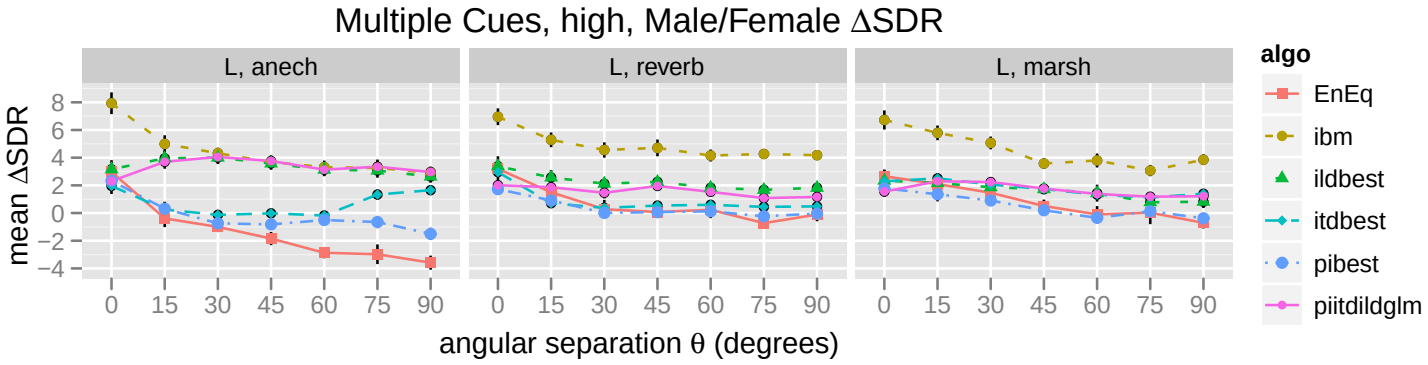


Figure 5.10: Linear supervised separation by multiple cues, high

5.7 Benefit of Computing Pitch per $t\nu$ Pixel

Figure 5.11 demonstrates the benefit of determining pitch per $t\nu$ pixel, instead of computing a pitch value per time instant. **pibestPraat** is the performance of best linear separation using pitch values from PRAAT as opposed to **bestpi** that gives best linear separation using pitch values per $t\nu$ pixel.

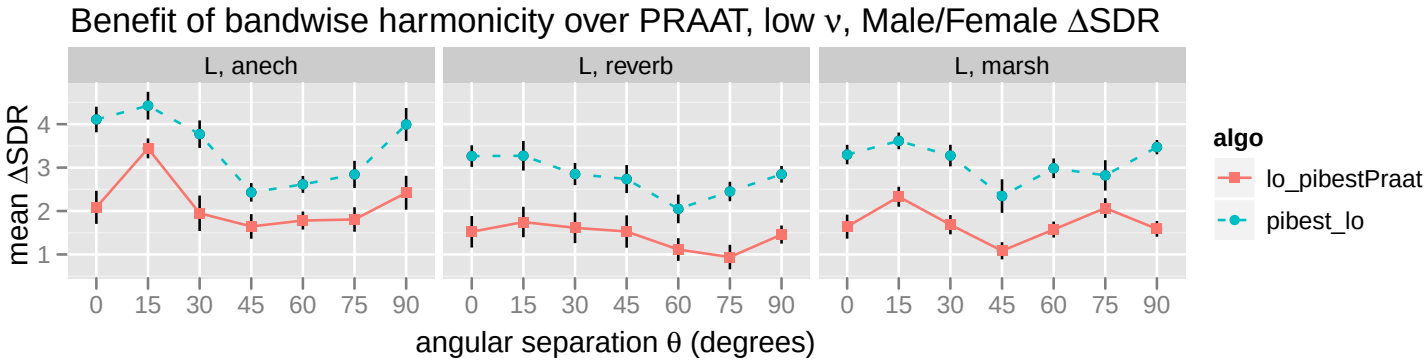


Figure 5.11: Benefit of bandwise harmonicity over PRAAT in low frequency bands

5.8 Baselines for Framewise Separation

Sequential segmentation of $t\nu$ pixels creates clusters for grouping, and the performance of the final the segmentation-grouping algorithm is bounded by the ideal binary segregation of segments into target and masker. Figure 5.12 compares ideal assignment of segments to best linear separation with supervised learning with multiple features (**piitdildglm**).

Ideal binary separation of segments is performed by attributing each segment to the source that contributes more than half the energy of constituting pixels according to the ideal binary mask.

- **seg0IBM** is the segmentation performed with no weights.
- **segHCIBM** is the segmentation performed using harmonicity and binaural coherence as reliability maps for pitch and ITD respectively.
- No difference in performance is observable between **seg0IBM** and **segHCIBM**, even though the latter algorithm was expected to be better. Reliability maps of features are used by **segHCIBM** to generate feature reliabilities for each segment that can be consumed at later stages of the algorithm. The performance difference between the two algorithms is not pursued further, and the segmentation used for **segHCIBM** is developed.

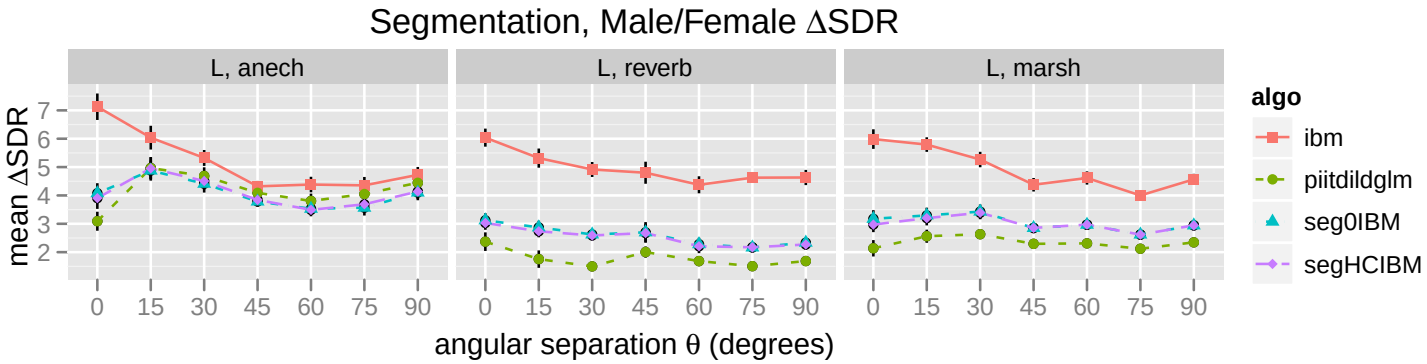


Figure 5.12: Segmentation baseline

5.9 Linear Separation of Segments

Figure 5.13 demonstrates best linear separation of segments by ITD alone (**seg0Lin** and **segHCLin** for segmentation with and without reliability maps respectively). The performance of linear segment separation is comparable to **piitdildglm** for most cases except anechoic, which is plausible since ILD, a major cue for anechoic situations is not used for this task.

Using multiple features for linear separation could potentially help raise performance above **piitdildglm**. However, instead of exploring this further, the fully unsupervised the segmentation-grouping algorithm is developed next.

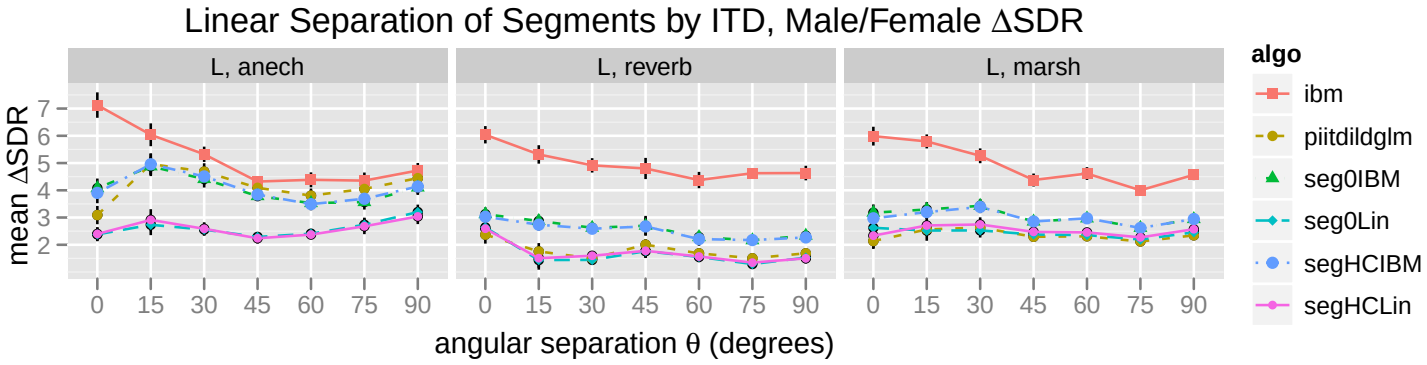


Figure 5.13: Segmentation followed by linear separation by ITD

5.10 Performance of The Segmentation-Grouping Algorithm

- **steppeSpace** is the segmentation and grouping algorithm run with spatial cues only, i.e. ITD and ILD only.
- **twotierHCEn** is the segmentation and grouing algorithm run with all available cues, i.e. pitch, ITD, ILD, center frequencies of spectral bands, and energy content per $t\nu$ pixel. Note that the performance of the algorithm at 0° is better than **steppeSpace**.

The Segmentation-Grouping Algorithm outperforms a best linear supervised algorithm in reverberant situations, but is outclassed in anechoic situation. A potential reason for this is a failure to effectively exploit ILD cues that are highly informative in anechoic situations. ILD is weighted at most as much any other highly *reliable* cue, regardless of whether that cue is *informative*.

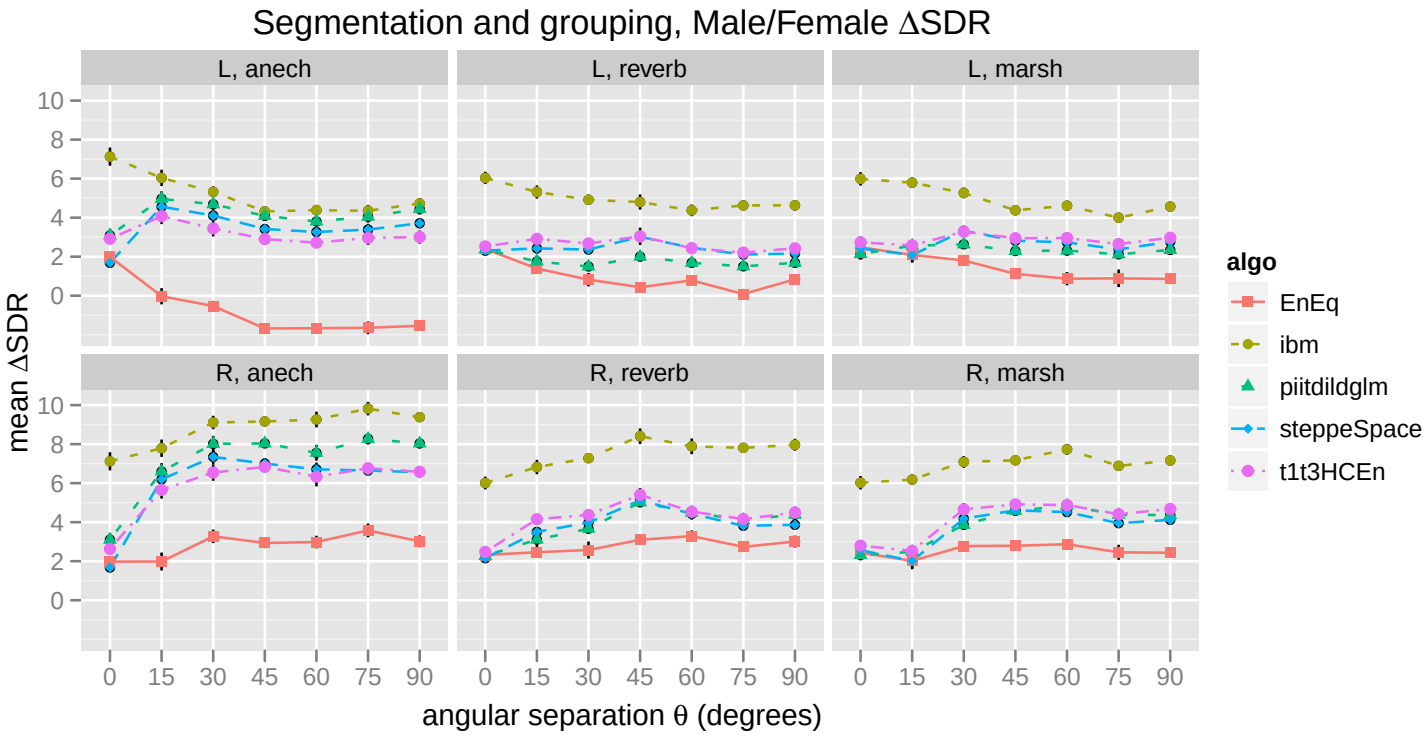


Figure 5.14: Segmentation followed by grouping

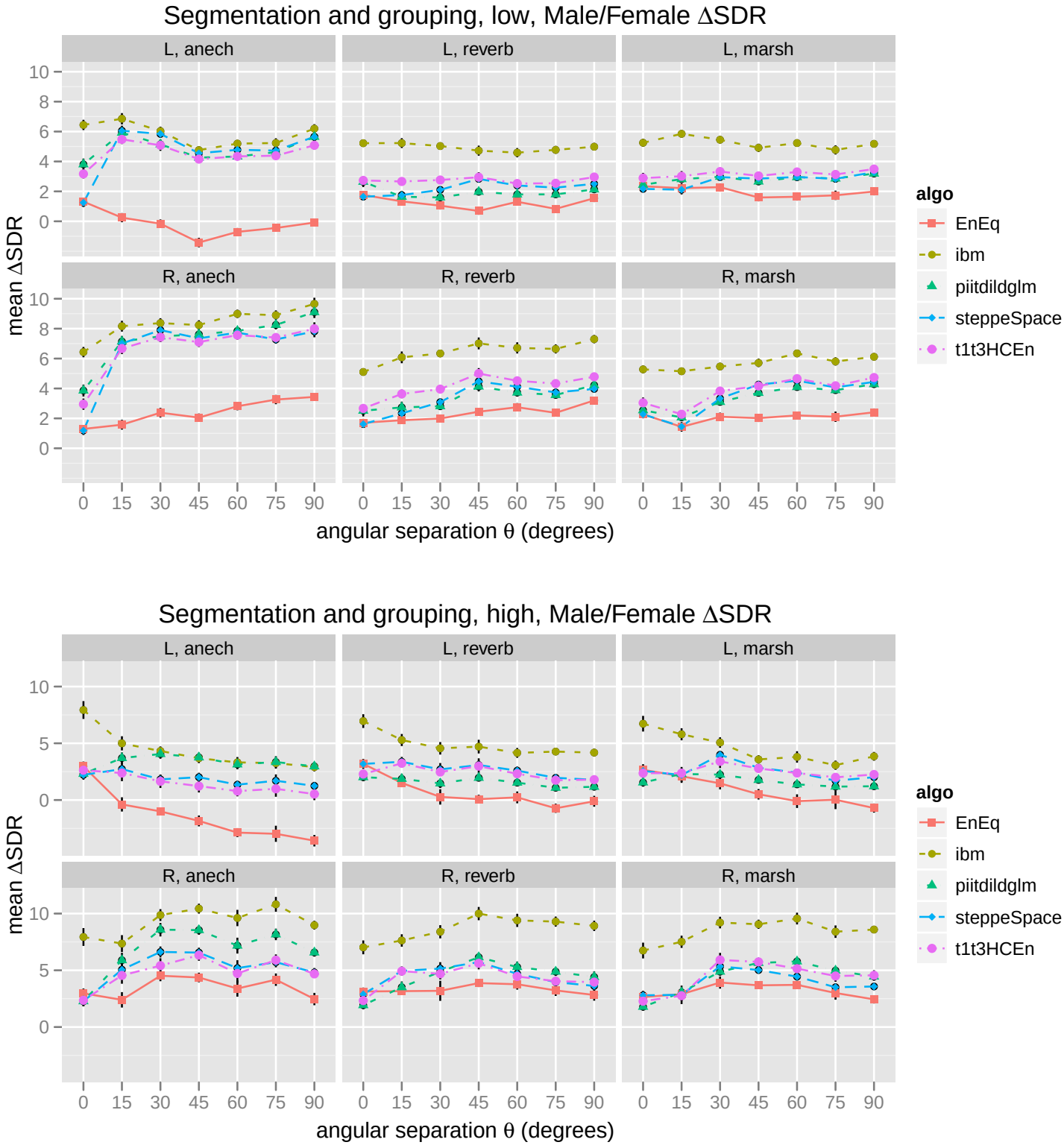


Figure 5.15: From top to bottom: (a) Segmentation followed by grouping, low (b) Segmentation followed by grouping, hi

5.11 Segmentation-Grouping versus Monolithic Grouping

Figure 5.16 shows **steppeUr**, an algorithm directly clusters the whole $t\nu$ map into clusters, with each cluster representing a separate speech source. Note that the performance of this algorithm suffers in reverberation as compared to **twotierHCEn**, since it does not have a mechanism to tie evidence from lower and higher frequency bands for speech separation.

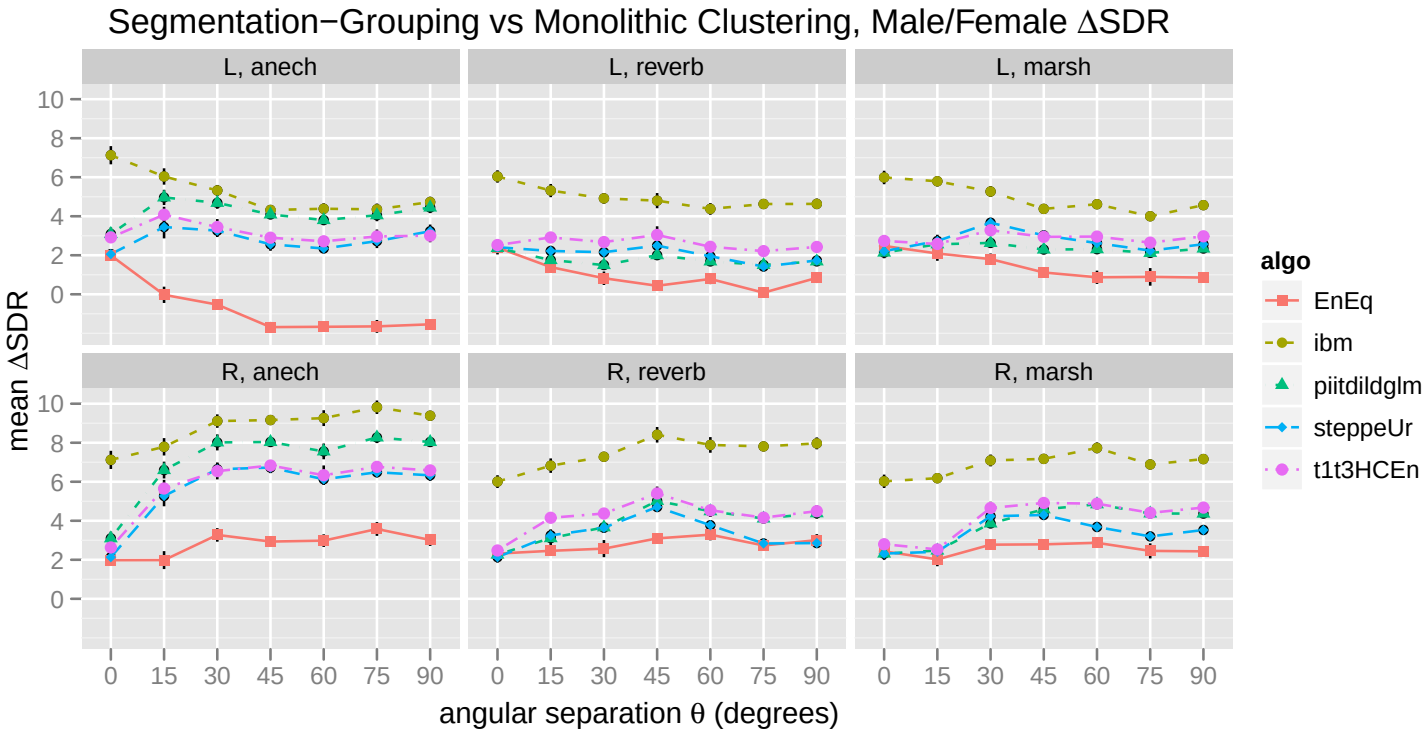


Figure 5.16: Sequential segmentation compared to monolithic grouping

5.12 Benefit of Reliability

Figure 5.17 shows **steppeSeg0**, an algorithm that performs segmentation without using feature reliabilities. While the benefit of using reliabilities in **twotierHCEn** over **steppeSeg0** is marginal, it does appear in the more difficult reverberant situations.

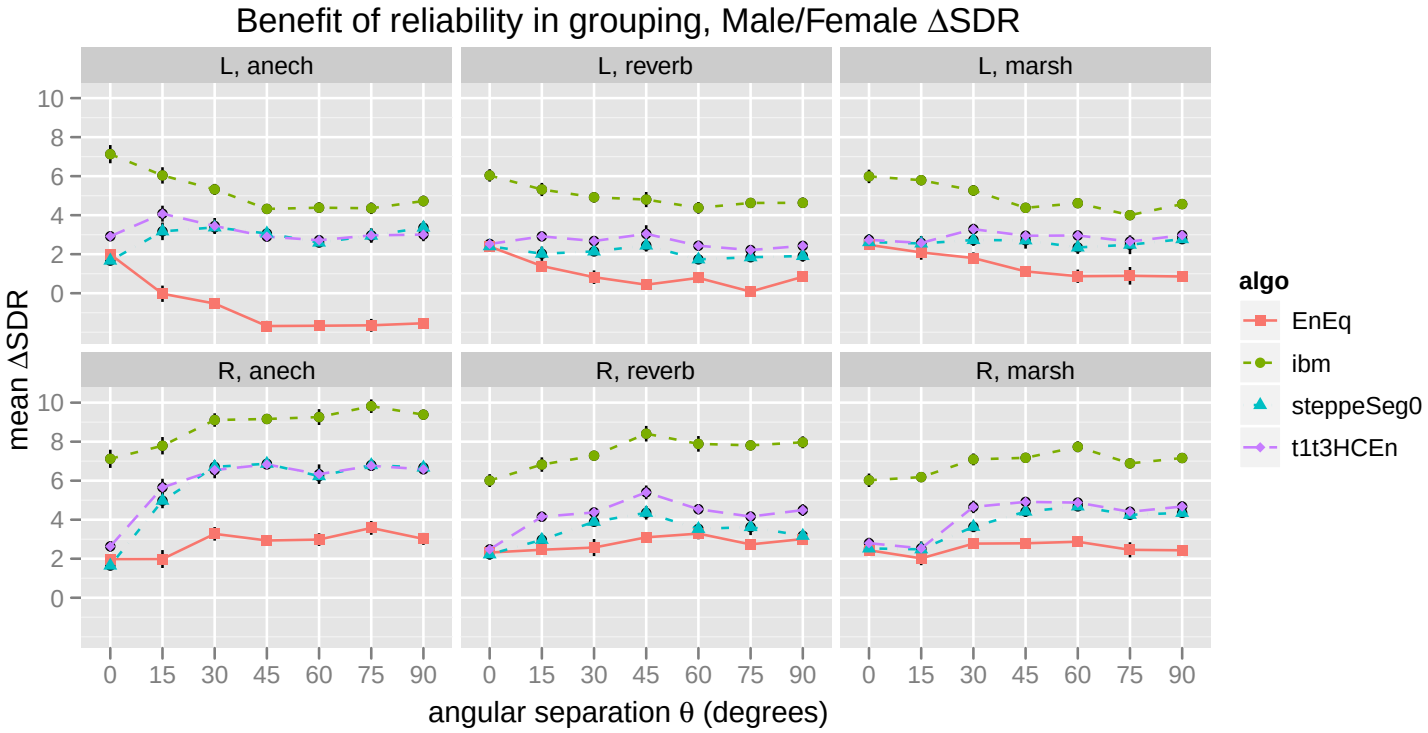


Figure 5.17: Benefit of reliability

5.13 Two talker mixture with both speakers of the same gender

The use of pitch cues in **steppeSegHarCohEn** does not benefit separation. This goes against the expectation that local pitch excursions might be helpful in separation even though the long term average pitch of the talkers may be similar.

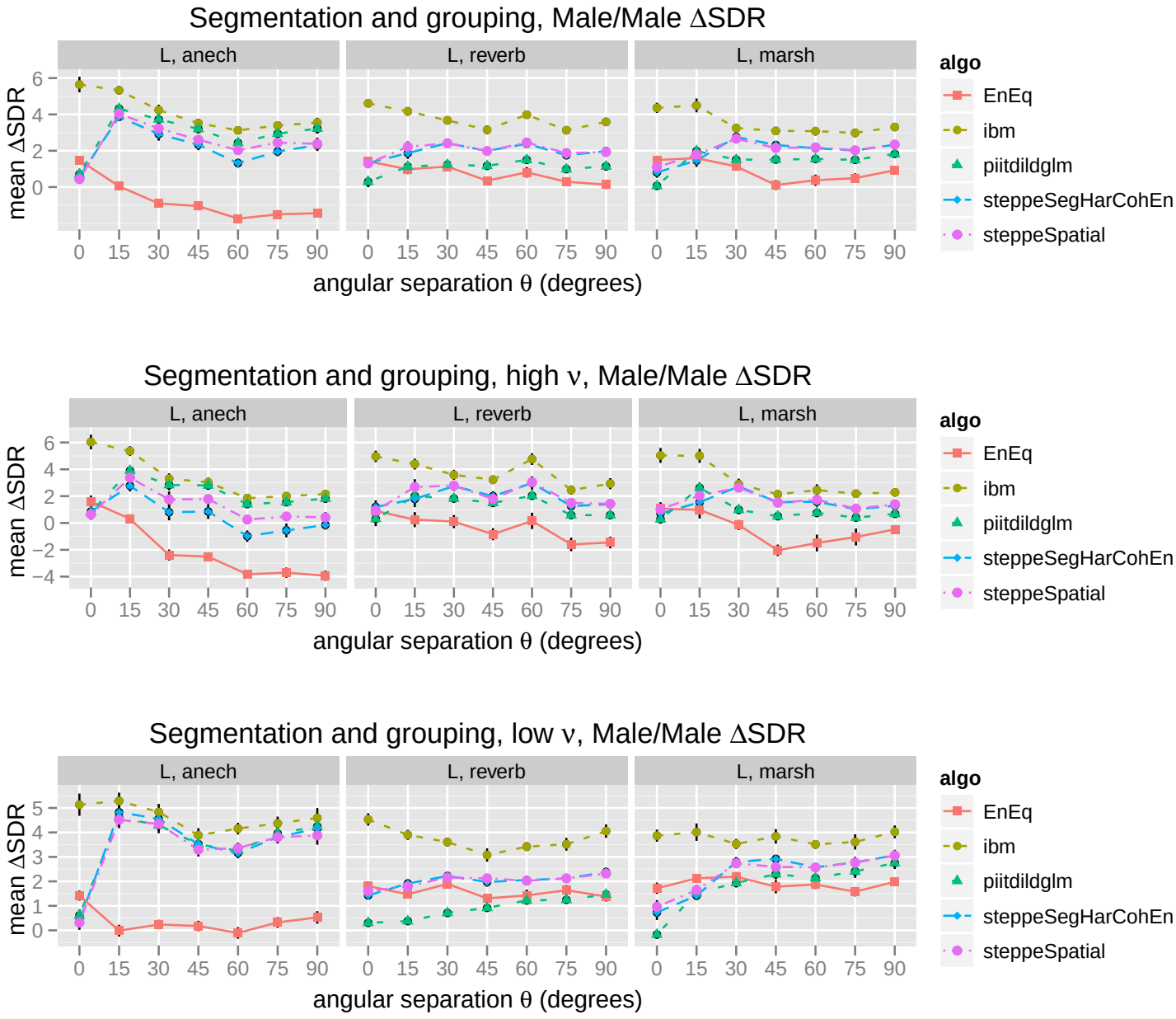


Figure 5.18: From top to bottom: (a) mm-segmentation-grp (b) mm-segmentation-grp-hi (c) mm-segmentation-grp-lo

5.14 Automatic Speech Recognition Evaluation

Automatic Speech Recognition (ASR) task is set up using PASCAL (2007). The 118 speech mixtures of two talkers reciting WSJ sentences are available from two six-channel microphone arrays. A trained HTK speech recognition engine is made available with the evaluation dataset, and is adapted to 64-band gammatone analysis–synthesized input to minimize impact of artifacts due to this processing.

Two channels are picked from one of six–channel inputs available. If channels are picked from different microphone arrays, the ITDs received are higher than a millisecond, and all other feature maps are not synchronized with the ITD values.

Each pair of channels is tried out from amongst the six channels, and ITDs are computed in a coarsely decomposed bandwise ITD is computed. The pair that has the highest spread of ITD values (entropy) is selected, since it is known that the mixture has two sources, and noise characteristics of the microphone are assumed the same.

Results are presented below, followed by some possible reasons for their not being stellar. .

5.14.1 Results

Description	Best	Separated
Number of words	1874	2071
%Word correct	60.03	8.11
%Accuracy	55.50	6.37
Number of sentences	118	118
%Sentence Correct	5.93	0.00

Table 5.2: Summary of ASR Results

5.14.2 Issues with ASR Evaluation

- *Lack of ILD* Microphone recordings do not have significant ILDs to help in high frequency separation.
- *Shorter ITDs* Since the recordings are from an array, and not a human head, the range of ITDs obtained lies within -500 to 500 μs.
- *Reverberation* Echoes are not removed from the mix. In the SDR measure, echoes belonging to the source are not penalized if they are correctly attributed to the source.
- *Distortions* Beyond smoothing the output mask, no significant effort is made to eliminate distortions from the output signal. ASR is highly sensitive to distortions, which make it suffer very rapid degradation in performance.

- *Source imbalance in mixtures* Speech mixtures are highly imbalanced, so that in several cases very short signal is combined with a long competing signal. While this is not an issue for the segmentation-grouping algorithm, the imbalance leads to loss of gains due to sequential segmentation where this algorithm may excel.

Chapter 6

Conclusion

6.1 Contribution of this effort

- *Multiple cues with reliabilities treated homogeneously.* The Segmentation-Grouping Algorithm demonstrates speech separation by computational auditory speech analysis without creating handcrafted rules for each feature. This is the first algorithm to combine multiple cues in a statistical framework so that they can be treated homogeneously. This is also the first algorithm to consider spectro–temporal maps of per–feature reliability for differential weighting of the cues, and demonstrates that employing cues is beneficial in the more challenging reverberant situations.
- *Unsupervised binaural algorithm.* The Segmentation-Grouping Algorithm does not require training data for performing the separation task, and relies instead on the inherent separability of the mixture in a sub–space of the features. Furthermore, the algorithm relies on two channels for spatial cues for separating any number of sources; and can operate with a single channel albeit with sub-par performance. This is in contrast to ICA that requires the same number of input channels as the number of sources to be unmixed.
- *Effectiveness in reverberation.* The Segmentation-Grouping Algorithm is demonstrated to be better than a supervised overfitted logistic regression model in reverberant situations. The integration of multiple features, each weighted by its reliability, at multiple time resolutions makes the separation more robust to reverberation than logistic regression, or monolithic clustering. Furthermore, this algorithm builds upon the features and competing algorithms that improve feature estimates are expected to further improve its performance.
- *Clarifies challenges.* The separation of evaluation into high and low frequency bands is an important step in clarifying the performance of speech separation algorithms. The effectiveness of a pitch algorithm cannot be judged without examining the separation performance during harmonic instants; and the performance of an ITD algorithm requires monitoring low frequency bands. While this may be obvious to a practitioner, results are not presented split by frequency bands in the literature.
- *Computation of pitch and ITD.* Pitch, or more accurately, auto–correlation peak lying between 50–400 Hz, is computed per $t\nu$ pixel, instead of per instant, enabling use of pitch as a feature. The correlation based algorithms for pitch (and harmonicity) and ITD (and coherence) are also almost identical. This is another point not emphasized in extant literature.

6.2 Limitations of The Segmentation-Grouping Algorithm and Extensions

- *Unsupervised algorithm.* Even though this is a benefit when no training data is available, or when training is not desirable, it is a severe limitation since the algorithm does not make use of available training data to improve, or adapt to a specific room or talker configuration.
- *Not a probabilistic model.* The ad-hoc treatment of cues limits the ability to exploit more features since parameters such as scaling and reliability weights are not derived in a principled manner. A supervised probabilistic formulation may be created based on the ideas discussed here; or a parameterized probabilistic formulation may be created with reasonable priors that can adapt parameters based on e.g. the first few seconds of a new room situation, or initial talker data.
- *Esoteric Approach.* The algorithm originates from a desire to exploit CASA, and misses out on several interesting speech separation applications. In anechoic conditions, the algorithm performance is not expected to match algorithms such as BSS. In reverberation, the algorithm does not effectively suppress ambient noise and reverberation – it attempts to classify the reverberant energy correctly as target or masker instead of eliminating it, and fails in ASR tasks.
- *Vertical integration with ASR is not straightforward.* Industrial ASR systems are vertically integrated, so that speech enhancement of a speech snippet is not performed in isolation of the recognition system. Integration into an ASR system has not be thought out for the current system.
- *Does not model higher order statistical information.* It would be difficult to effectively and efficiently model pitch tracks, frequency glides, spectro–temporal patterns etc. as $t\nu$ maps.
- *Prone to errors in case of noisy or badly scaled data.* The measure of information in a feature for separation is not available; and reliability is used instead. This limits the extent to which irrelevant features can be discarded, e.g. pitch in same gender talker mixture, spatial cues from a mixture with co–located sources, etc. Because of this, the algorithm performance is observed to be lower when pitch is used in *male*/male conditions; and lower than **bestild** in anechoic situations.
- *Handcrafted features.* Features are treated homogeneously only after feature map creation, which requires manually crafting the feature under the constraints of feature linearity across time and frequency. It is not always clear how a feature may be utilized, e.g. pitch can be represented both as frequency, or as a period. Deriving reliability maps is similarly challenging. In the current work, reliability maps were validated by performance on logistic regression models using the proposed reliability as feature weighting for linear separation.
- *Underutilized onsets and offsets.* The algorithm does not show any benefit of using energy differentials. Even worse, it does not have a clear way to rule out deriving benefit from these features.
- *It is a linear separator.* After the initial framewise segmentation, the grouping stage creates linear boundaries between talkers in feature space. Non–linear separation with a data driven technique like

spectral clustering (Bach and Jordan, 2009) would offer better separation and guide feature selection and feature weighting.

- *Domain specific algorithm.* The algorithm is coarse, and not expected to work well on other tasks such as musical signal separation, or mixed neural data separation, or resolution of time series into component biomarker data.
- *No realistic evaluation setup.* The speech separation problem requires a standard evaluation metric and benchmark tasks that isolate various components of the task – dereverberation, denoising, environment noise reduction, phoneme reconstruction, speaker identification, and automatic speech recognition.

In conclusion, the segmentation-grouping algorithm is a prototype that demonstrates some interesting ideas of CASA, but is not mature enough to conclusively confirm or refute any.

Appendix A

Evaluation of The Segmentation-Grouping Algorithm on three-talker mixtures

Three talker mixture is denoted as, for example, *Male*/Male/Male, indicating that the emphasized male talker is the target. Spatially, the middle talker is located right ahead, with the flanking talkers at either side separated by azimuthal angle that is plotted along the x–axis in the figures. Since the first talker is always the target, left ear is the better ear, and the only one plotted here.

Salient points in three talker mixture performance are:

- Performance is better than **EnEq**in most situations, except in *Male*/Male/Male and *Male*/Female/Male cases in classroom. In these relatively tougher situations, the benefit of spatial separation starts showing after a separation of more than 30°.
- Performance when target and maskers are co–located is near **EnEq**in all cases. The clustering algorithms at segmentation and grouping stages are swamped by the noise from multiple features, and are not able to exploit pitch.

A.1 Three talker mixture FMM

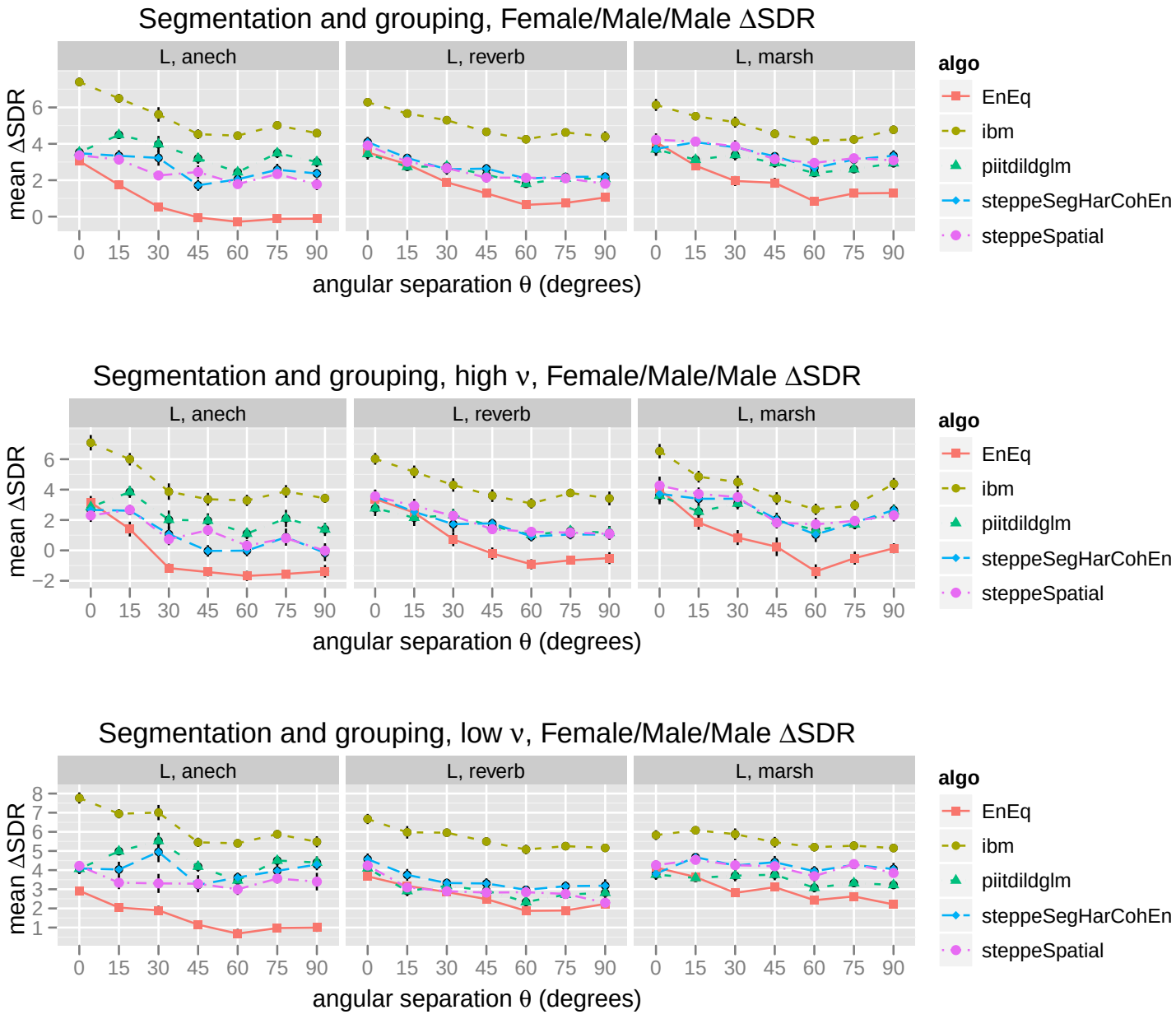


Figure A.1: From top to bottom: (a) fmm-segmentation-grp (b) fmm-segmentation-grp-hi (c) fmm-segmentation-grp-lo

A.2 Three talker mixture MMM

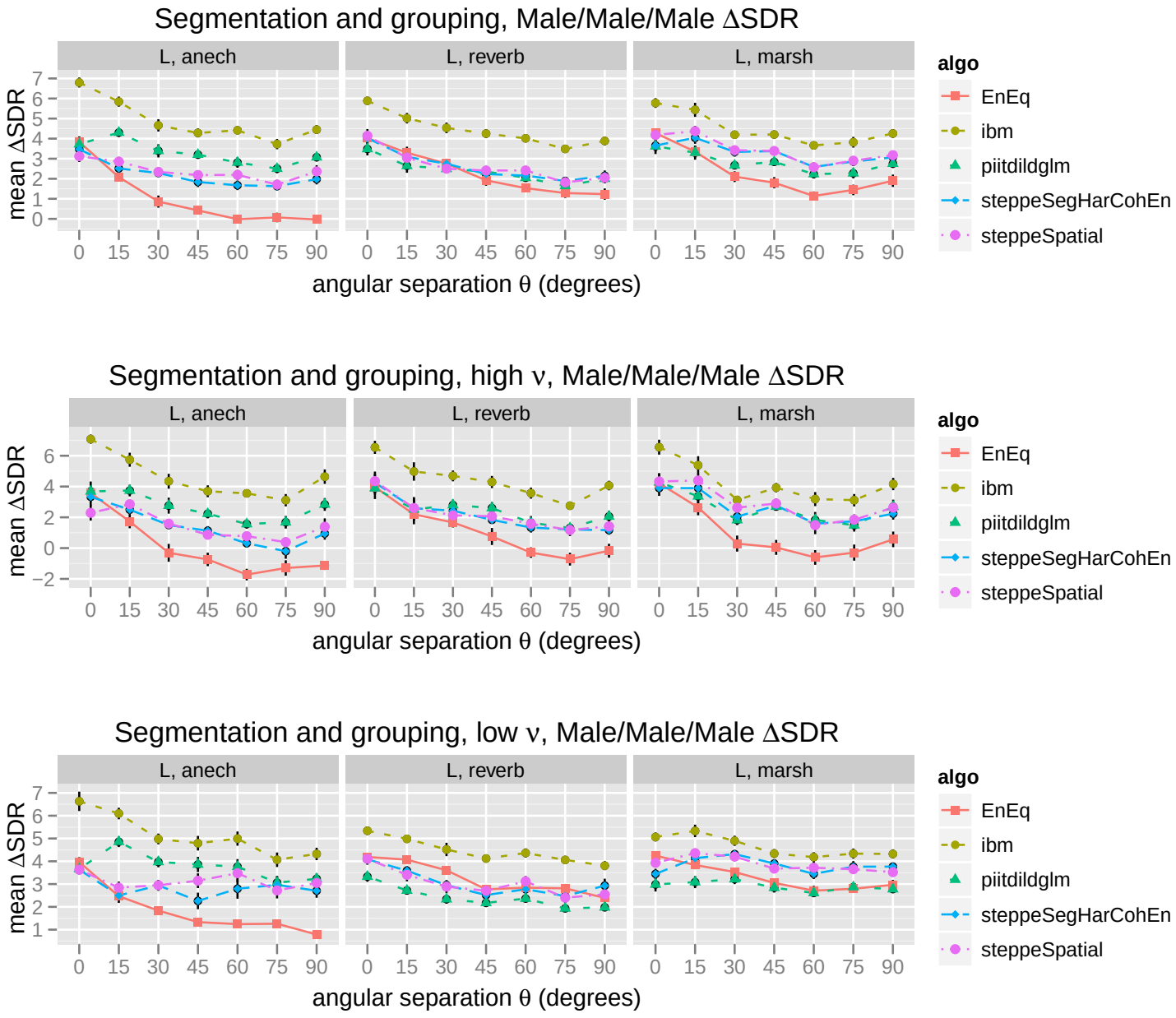


Figure A.2: From top to bottom: (a) mmm-segmentation-grp (b) mmm-segmentation-grp-hi (c) mmm-segmentation-grp-lo

A.3 Three talker mixture MFM

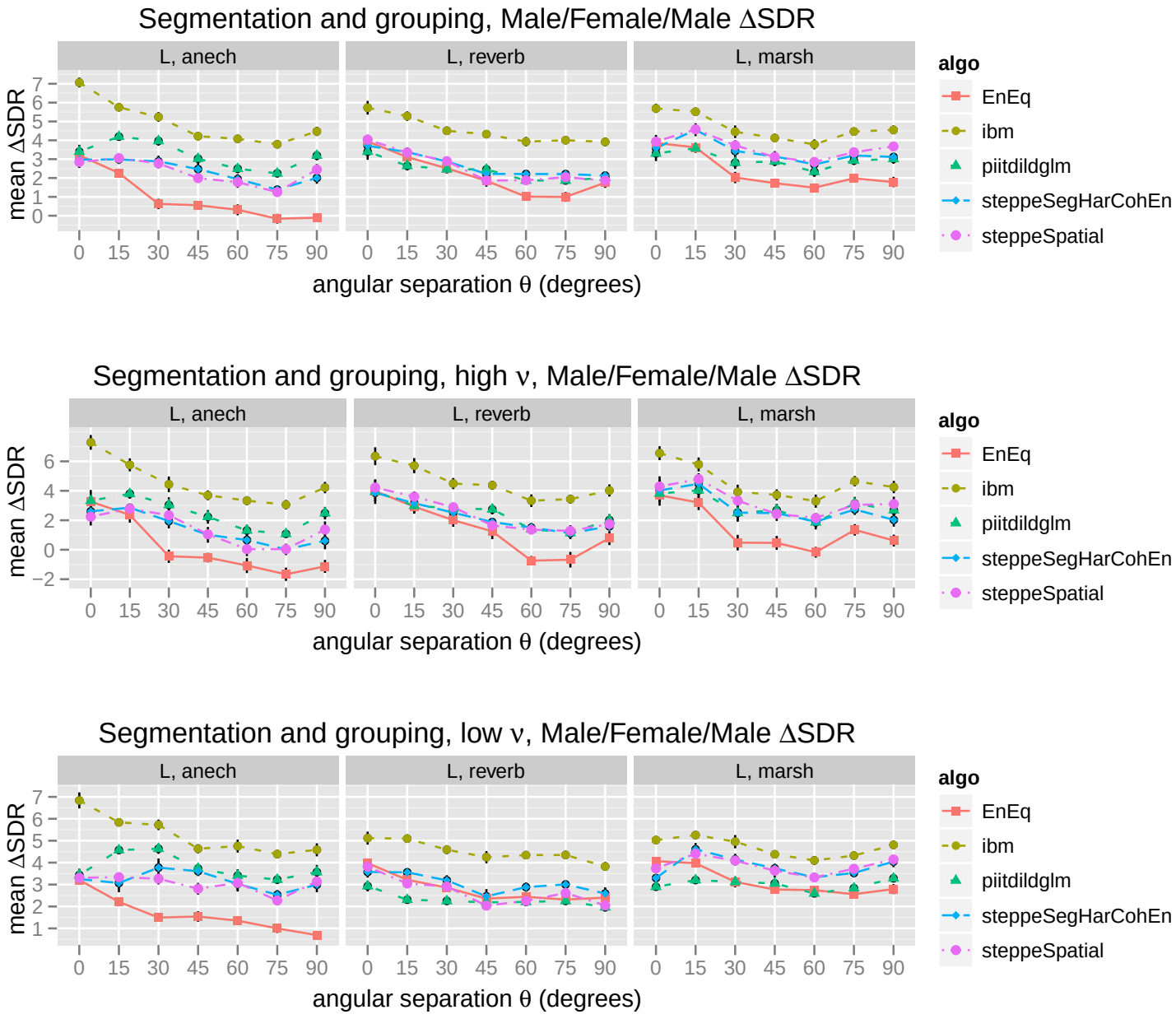


Figure A.3: From top to bottom: (a) mfm-segmentation-grp (b) mfm-segmentation-grp-hi (c) mfm-segmentation-grp-lo

Appendix B

Correlation based algorithms for Pitch and ITD

Algorithm 5: Autocorrelation peak algorithm for Pitch and Harmonicity

Data: Monaural signal $S(\nu, t)$
Result: Autocorrelation peaks $\Pi(\nu, t)$
Harmonicity $H(\nu, t)$

1. *Normalized per-band autocorrelation*
$$A(\nu, t, \tau) = \left[\sum_{t_i=t-\delta}^{t+\delta} S'(t_i - \frac{\tau}{2}) S'(t_i + \frac{\tau}{2}) \right] / \left[\sum_{t_i=t-\delta}^{t+\delta} S'^2(\nu, t_i) \right]$$
where windowed signal $S'(\nu, t) = S(\nu, t).w(t)$.
2. *Summary autocorrelation*
$$A_\Sigma(t, \tau) = \sum_\nu A(\nu, t, \tau)$$
where τ covers the domain of plausible pitch frequencies.
3. *Candidate pitch detection by thresholding summary autocorrelation*
$$A_\theta(t, \tau) = \begin{cases} 1 & \text{if } A_\Sigma(t, \tau) > \theta_t, \\ 0 & \text{if } A_\Sigma(t, \tau) \leq \theta_t \end{cases}$$
where θ_t is the value at the kth percentile of $A_\Sigma(t, \tau)$ distributed over τ.
4. *Per-band autocorrelation emphasized by candidate pitch mask*
$$A_e(\nu, t, \tau) = A_\theta(t, \tau).A(\nu, t, \tau)$$
5. *Pitch is the frequency corresponding to the peak location*
$$\Pi(\nu, t) = \frac{1}{\arg_\tau \max A_e(\nu, t, \tau)}$$
6. *Harmonicity is the peak height*
$$H(\nu, t) = \max A_e(\nu, t, \tau)$$

Algorithm 6: Cross–correlation peak algorithm for ITD and Coherence

Data: Binaural signal $S(c, \nu, t)$
Result: Interaural time difference $X(\nu, t)$
Coherence $H(\nu, t)$

1. *Normalized per-band cross–correlation*

$$X(\nu, t, \tau) = \left[\sum_{t_i=t-\delta}^{t+\delta} S'(l, t_i - \frac{\tau}{2}) S'(r, t_i + \frac{\tau}{2}) \right] / \left[\sqrt{\sum_{t_i=t-\delta}^{t+\delta} S'^2(l, \nu, t_i) \sum_{t_i=t-\delta}^{t+\delta} S'^2(r, \nu, t_i)} \right]$$

where windowed signal $S'(c, \nu, t) = S(c, \nu, t).w(t)$.

2. *Summary cross–correlation*

$$X_\Sigma(t, \tau) = \sum_\nu X(\nu, t, \tau)$$

where τ covers the domain of plausible interaural time difference.

3. *Per-band cross–correlation emphasized by summary cross–correlation*

$$X_e(\nu, t, \tau) = X_\theta(t, \tau).X(\nu, t, \tau)$$

4. *ITD corresponds to the peak location*

$$T(\nu, t) = \arg_\tau \max X_e(\nu, t, \tau)$$

5. *Coherence is the peak height*

$$H(\nu, t) = \max X_e(\nu, t, \tau)$$

Appendix C

Verbatim Results of Automatic Speech Recognition Task

Results on Separated Speakers

```
------------------------------ Speaker Results ------------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
16_17:   6.82(  6.29)  [H=  39, D=211, S=322, I=  3, N= 572]   0.00 [N= 31]
17_18:  11.64( 10.40)  [H=  56, D=157, S=268, I=  6, N= 481]   0.00 [N= 28]
18_19:   9.22(  6.07)  [H=  38, D=159, S=215, I= 13, N= 412]   0.00 [N= 23]
19_20:   5.81(  3.03)  [H=  23, D=183, S=190, I= 11, N= 396]   0.00 [N= 22]
20_16:   5.71(  4.29)  [H=  12, D= 91, S=107, I=  3, N= 210]   0.00 [N= 14]
------------------------------ Overall Results ------------------------------
SENT: %Correct=0.00 [H=0, S=118, N=118]
WORD: %Corr=8.11, Acc=6.37 [H=168, D=801, S=1102, I=36, N=2071]
=============================================================================
```

Results on Original Speakers before Mixing (Lapel)

```
------------------------------ Speaker Results ------------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
16_17:  47.08( 43.18)  [H= 145, D= 47, S=116, I= 12, N= 308]   0.00 [N= 19]
17_18:  48.96( 42.75)  [H= 189, D= 44, S=153, I= 24, N= 386]   0.00 [N= 23]
18_19:  73.25( 69.00)  [H= 241, D= 25, S= 63, I= 14, N= 329]  13.64 [N= 22]
19_20:  72.97( 69.14)  [H= 305, D= 28, S= 85, I= 16, N= 418]  11.11 [N= 27]
20_16:  56.58( 52.19)  [H= 245, D= 54, S=134, I= 19, N= 433]   3.70 [N= 27]
------------------------------ Overall Results ------------------------------
SENT: %Correct=5.93 [H=7, S=111, N=118]
WORD: %Corr=60.03, Acc=55.50 [H=1125, D=198, S=551, I=85, N=1874]
=============================================================================
```

Table C.1: ASR verbatim results

References

i2002sld Aarabi, P. Self-localizing dynamic microphone arrays. *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 32(4): 474–484, 2002. (Cited on pages 11 and 60.)

r2002tdb Aichner, R., S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari. Time-domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming. In *Proc. Neural Networks for Signal Processing*, pages 445–454, 2002. (Cited on pages 5 and 11.)

modeling Assmann, P. and Q. Summerfield. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 1990. (Cited on page 45.)

spectral Bach, F. and M.I. Jordan. Spectral clustering for speech separation. *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, page 221, 2009. (Cited on pages 4 and 79.)

n2003bbc Baumann, W., D. Kolossa, and R. Orglmeister. Beamforming-based convolutive source separation. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 5, 2003. (Cited on pages 5 and 11.)

l1995ima Bell, A. and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. (Cited on pages 5 and 11.)

:bregman Bregman, A. S. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990. (Cited on pages 8, 9, 13, and 14.)

orst2000 Bronkhorst, A. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica*, 86:117–128, 2000. (Cited on pages 13 and 19.)

n1994cas Brown, G. J. and Martin Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994. (Cited on pages 5, 12, 14, 16, 18, 21, 24, and 38.)

t2005bmt Brungart, D., B.D. Simpson, P.S. Chang, and D.L. Wang. A binary masking technique for isolating energetic masking in speech perception. *The Journal of the Acoustical Society of America*, 117:2484, 2005. (Cited on page 14.)

uss2003b Buss, E., Joseph W 3rd Hall, and John H Grose. The masking level difference for signals placed in masker envelope minima and maxima. *J Acoust Soc Am*, 114(3):1557–1564, Sep 2003. (Cited on page 14.)

oell1997 Campbell Jr, J. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997. (Cited on page 16.)

lspatial Cepeda, M., V. Best, and B. Shinn-Cunningham. Spatial influences on the spectral restoration of narrowband speech. *The Journal of the Acoustical Society of America*, 129:2589, 2011. (Cited on page 2.)

y1953ser Cherry, E. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25:975, 1953. (Cited on page 1.)

n1994ica Comon, P. Independent component analysis, a new concept. *Signal Processing*, 36(3):287–314, 1994. (Cited on pages 5 and 11.)

elle2001 Colburn, H. S. and S. K. Isabelle. Physiologically based models of binaural detection. *Physiological and Psychophysical Bases of Auditory Function*, pages 161–168, 2001. (Cited on page 17.)

ling1995 Culling, J. and Q. Summerfield. Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America*, 98:785, 1995. (Cited on page 18.)

rwin1997 Darwin, C. J. Auditory grouping. *Trends in Cognitive Science*, pages 327–333, 1997. (Cited on pages 14, 15, and 18.)

rwin1995 Darwin, C. J. and R P Carlyon. *Auditory Grouping*. Academic Press, 1995. (Cited on pages 8, 14, and 15.)

n1986mhv Darwin, C. and RB Gardner. Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *The Journal of the Acoustical Society of America*, 79:838, 1986. (Cited on page 9.)

e1993sch de Cheveigné, A. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *The Journal of the Acoustical Society of America*, 93:3271, 1993. (Cited on page 12.)

e1998cmp de Cheveigné, A. Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, 103:1261, 1998. (Cited on page 16.)

n1992pea Denbigh, P. N. and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech communication*, 11(2-3):119–125, 1992. (Cited on page 10.)

oata2003 Ebata, M. Spatial unmasking and attention related to the cocktail party problem. *Acoustical Science and Technology*, 24:208–219, 2003. (Cited on page 19.)

s1996pdc Ellis, D. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996. (Cited on pages 5 and 12.)

llis:ess Ellis, D. Evaluating speech separation systems. *Speech Separation by Humans and Machines*, pages 295–304. (Cited on page 22.)

wley2004 Hawley, M. L., Ruth Y. Litovsky, and John F. Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843, 2004. URL http://link.aip.org/link/?JAS/115/833/1. (Cited on pages 14, 17, 18, and 19.)

ykin2005 Haykin, S. and Zhe Chen. The cocktail party problem. *Neural Computation*, 17:1875–1902, 2005. (Cited on page 10.)

ustering Himmelspach, L. and S. Conrad. Clustering approaches for data with missing values: Comparison and evaluation. In *Digital Information Management (ICDIM), 2010 Fifth International Conference on*, pages 19–28. IEEE. (Cited on page 32.)

u2002mss Hu, G. and D.L. Wang. Monaural speech separation. *Proceedings of Neural Information Processing Systems*, 2002. (Cited on pages 12 and 24.)

wang2004 Hu, G. and D.L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *Neural Networks, IEEE Transactions on*, 15(5):1135–1150, 2004. (Cited on pages 4, 14, 16, 21, and 24.)

wang2007 Hu, G. and D.L. Wang. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:396–405, 2007. (Cited on page 22.)

n2000ica Hyvärinen, A. and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. (Cited on pages 5 and 11.)

d2005rrr Kidd, G., C.R. Mason, A. Brughera, and W.M. Hartmann. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acustica united with Acta Acustica*, 2005a. (Cited on page 21.)

kidd2005 Kidd, G. J., Tanya L Arbogast, Christine R Mason, and Frederick J Gallun. The advantage of knowing where to listen. *J Acoust Soc Am*, 118(6):3804–15, 2005b. (Cited on pages 18 and 19.)

rberated Kim, L., I. Tashev, and A. Acero. Reverberated speech signal separation based on regularized subband feedforward ica and instantaneous direction of arrival. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2678–2681. IEEE, 2010. (Cited on page 5.)

n2004sms Kristjansson, T., H. Attias, and J. Hershey. Single microphone source separation using high resolution signal reconstruction. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 2, 2004. (Cited on page 11.)

e1997bsd Lee, T., A.J. Bell, and R.H. Lambert. Blind Separation of Delayed and Convolved Sources. *Proceedings of Neural Information Processing Systems*, pages 758–764, 1997. (Cited on pages 5 and 11.)

li:1673 Li, N. and Philipos C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *The Journal of the Acoustical Society of America*, 123(3):1673–1682, 2008. doi: 10.1121/1.2832617. URL http://link.aip.org/link/?JAS/123/1673/1. (Cited on page 30.)

timality Li, Y. and D.L. Wang. On the optimality of ideal binary time-frequency masks. *Speech Communication*, 51(3):230–239, 2009. (Cited on page 30.)

o1994sss Luo, H. and PN Denbigh. A speech separation system that is robust to reverberation. In *Speech, Image Processing and Neural Networks, 1994. Proceedings, ISSIPNN'94., 1994 International Symposium on*, pages 339–342, 1994. (Cited on page 10.)

lyon1996 Lyon, R. F. Re: Gammatone analysis/synthesis. *Auditory.org Mailing List*, 26 Sep 1996. URL http://www.auditory.org/mhonarc/1996/msg00127.html. (Cited on page 30.)

009ideal Mandel, M. and D. Ellis. The ideal interaural parameter mask: A bound on binaural separation systems. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 85–88. IEEE. (Cited on page 25.)

l2007alm Mandel, M., D.P.W. Ellis, and T. Jebara. An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments. *Proceedings of Neural Information Processing Systems*, 19:953, 2007. (Cited on pages 11 and 22.)

aluating Mandel, M., S. Bressler, B. Shinn-Cunningham, and D.P.W. Ellis. Evaluating source separation algorithms with reverberant speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1872–1883, 2010a. (Cited on pages ix and 20.)

andel09a Mandel, M. I., Ron J. Weiss, and Daniel P. W. Ellis. Model-based expectation maximization source separation and localization. *IEEE Transactions on audio, speech, and language processing*, 18(2):382–394, February 2010b. doi: 10.1109/TASL.2009.2029711. URL http://mr-pc.org/work/taslp10.pdf. (Cited on page 5.)

r1991efa Mellinger, D. *Event Formation and Separation in Musical Sound*. PhD thesis, Stanford University, 1991. URL ftp://ccrma-ftp.stanford.edu/pub/Publications/Theses/DAVEMThesis.ps.Z. (Cited on pages 11 and 23.)

:moore97 Moore, B. C. J. *An Introduction to the Psychology of Hearing*. Academic Press, fourth edition, 1997. (Cited on pages 15 and 17.)

oore1999 Moore, B. C. J. Neurobiology. Modulation minimizes masking. *Nature*, 397(6715):108–9, 1999. (Cited on page 14.)

oore1987 Moore, D. Physiology of higher auditory system. *British medical bulletin*, 43(4):856, 1987. (Cited on page 8.)

nham2001 Oxenham, A. J. and Torsten Dau. Modulation detection interference: Effects of concurrent and sequential streaming. *The Journal of the Acoustical Society of America*, 110(1):402–408, 2001. URL http://link.aip.org/link/?JAS/110/402/1. (Cited on page 14.)

s1976ssi Parsons, T. Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60:911, 1976. (Cited on page 10.)

hallenge PASCAL. Speech separation challenge II. 2007. URL http://homepages.inf.ed.ac.uk/mlincol1/SSC2/. (Cited on pages 38 and 74.)

rson1992 Patterson, R. D., K Robinson, J Holdsworth, D McKeown, C Zhang, and M Allerhand. *Complex sounds and auditory images*, pages 429–446. Pergamon, Oxford, 2004. (Cited on page 28.)

g1997dbn Peissig, J. and B. Kollmeier. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *The Journal of the Acoustical Society of America*, 101:1660, 1997. (Cited on page 18.)

6effects Poissant, S., N.A. Whitmal III, and R.L. Freyman. Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *The Journal of the Acoustical Society of America*, 119:1606, 2006. (Cited on page 25.)

d2002awd Rickard, S., Z. Yilmaz, and S.C. Res. On the approximate W-disjoint orthogonality of speech. In *Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on*, volume 1, 2002. (Cited on page 14.)

oman2003 Roman, N., DeLiang Wang, and Guy J. Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003. URL http://link.aip.org/link/?JAS/114/2236/1. (Cited on pages 4, 12, 18, 21, and 24.)

i1990mam Saberi, K. and D.R. Perrott. Minimum audible movement angles as a function of sound source trajectory. *The Journal of the Acoustical Society of America*, 88:2639, 1990. (Cited on page 39.)

i2001ssa Shamsoddini, A. and PN Denbigh. A sound segregation algorithm for reverberant conditions. *Speech Communication*, 33(3):179–196, 2001. (Cited on page 10.)

nham2007 Shinn-Cunningham, B. G., Lee A. K. C., and A. J. Oxenham. Auditory non-allocation of a sound element lost in perceptual competition. *Proceedings of the National Academy of Sciences*, 104:12223–12227, 2007. (Cited on page 18.)

oarb2005 Shinn-Cunningham, B. Influences of spatial cues on grouping and understanding sound. *ForumAcusticum*, 2005. (Cited on pages 19 and 27.)

ntje2004 Shinn-Cunningham, B., Antje Ihlefeld, Satyavarta, and Eric Larson. Bottom–up and top–down influences on spatial unmasking. *Acta Acustica united with Acustica*, 91:967–979, 2005. (Cited on page 18.)

m2008oba Shinn-Cunningham, B. Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5):182–186, 2008. (Cited on page 21.)

ey1994at Slaney, M. Auditory toolbox. *Apple Computer, Inc. Technical Report*, 45, 1994. (Cited on page 29.)

y1998cpa Slaney, M. A critique of pure audition. *Computational Auditory Scene Analysis*, pages 27–42, 1998. (Cited on page 8.)

auditory Slaney, M., D. Naar, and R.F. Lyon. Auditory model inversion for sound separation. In *Proceedings of 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 77–80. Citeseer, 1994. (Cited on page 30.)

s2001rrc Smaragdis, P. *Redundancy Reduction for Computational Audition, a Unifying Approach*. PhD thesis, Massachusetts Institute of Technology, 2001. (Cited on pages 22 and 23.)

09sparse Smaragdis, P., M. Shashanka, and B. Raj. A sparse non-parametric approach for single channel separation of known sounds. In *Proc. Neural Information Processing Systems*, 2009. (Cited on page 5.)

s1990ass Stubbs, R. and Q. Summerfield. Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 87:359, 1990. (Cited on page 10.)

ield1992 Summerfield, Q. and John F. Culling. Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency. *The Journal of the Acoustical Society of America*, 92(4):2317–2317, 1992. URL http://link.aip.org/link/?JAS/92/2317/3. (Cited on page 15.)

a1998bss Torkkola, K. Blind signal separation in communications: Making use of known signal distributions. In *Proceedings of the 1998 IEEE Digital Signal Processing Workshop, Bryce Canyon, UT*, 1998. (Cited on pages 5 and 11.)

ouwe2001 van der Kouwe, A. J. W., DeLiang Wang, and Guy J Brown. A comparison of auditory and blind separation techniques for speech segregation. *IEEE Transactions on Speech and Audio Processing*, 9(3):189–195, 2001. (Cited on page 12.)

n1988bva Van Veen, B. and KM Buckley. Beamforming: a versatile approach to spatial filtering. *ASSP Magazine, IEEE [see also IEEE Signal Processing Magazine]*, 5(2):4–24, 1988. (Cited on pages 5 and 11.)

t2005pmb Vincent, E., R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Speech and Audio Proc*, 2005a. (Cited on pages 38 and 51.)

cent:bas Vincent, E., M.G. Jafari, S.A. Abdallah, MD Plumbley, and ME Davies. Blind audio source separation. Technical report, Tech. Rep. C4DM-TR-05-01, Centre for Digital Music, Queen Mary University of London, 2005b. (Cited on page 10.)

005ideal Wang, D. On ideal binary mask as the computational goal of auditory scene analysis. *Speech separation by humans and machines*, pages 181–197, 2005. (Cited on page 30.)

g1999ssi Wang, D. and GJ Brown. Separation of speech from interfering sounds based on oscillatory correlation. *Neural Networks, IEEE Transactions on*, 10(3):684–697, 1999. (Cited on pages 4 and 12.)

raub1986 Weintraub, M. A computational model for separating two simultaneous talkers. *Acoustics, Speech, and Signal processing, IEEE International Conference on*, 11:81–84, Apr 1986. (Cited on pages 4, 11, 14, 21, 23, and 24.)

u2003mta Wu, M., D.L. Wang, and G.J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11(3):229–241, 2003. (Cited on pages 12, 21, and 24.)

z2004bss Yilmaz, O. and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 52(7):1830–1847, 2004. (Cited on pages 4, 11, 24, and 60.)

:zhang01 Zhang, Heinz, Bruce, and Carney. A phenomenological model for the responses of auditory–nerve fibers: I. nonlinear timing with compression and suppression. *The Journal of Acoustic Society of America*, 109(2), February 2001. (Cited on page 29.)