10

Single-Microphone Speech Separation: The use of Speech Models

S. W. Lee
Singapore

1. Introduction

Separation of speech sources is fundamental for robust communication. In daily conversations, signals reaching our ears generally consist of target speech sources, interference signals from competing speakers and ambient noise. Take an example, talking with someone in a cocktail party and making a phone call in a train compartment. Fig. 1 shows a typical indoor environment having multiple sound sources, such as speech from different speakers, sounds from a television set and telephone ringing, etc. These sources are often overlapped in time and frequency. While human attends to individual sources without difficulty, most speech applications are vulnerable and resulted in degraded performance.

This chapter focuses on speech separation for single microphone input, in particular, the use of prior knowledge in the form of speech models. Speech separation for single microphone input refers to the estimation of individual speech sources from the mixture observation. It remains important and beneficial to various applications, namely surveillance systems, auditory prostheses, speech and speaker recognition.

Over the years, extensive effort has been devoted. Speech enhancement and separation are two popular approaches. Speech enhancement (Lim, 1983; Loizou, 2007) generally reduces the interference power, by assuming that certain characteristics of individual source signals are held. There is one speech source at most. In contrast, speech separation (Cichocki & Amari, 2002; van der Kouwe et al., 2001) extracts multiple target speech sources directly.

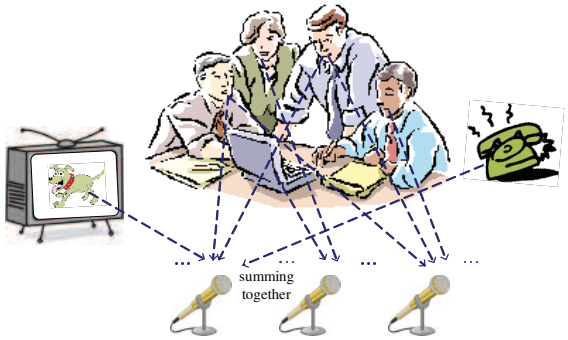


Fig. 1. Illustration of multiple sound sources present in typical acoustic environments.

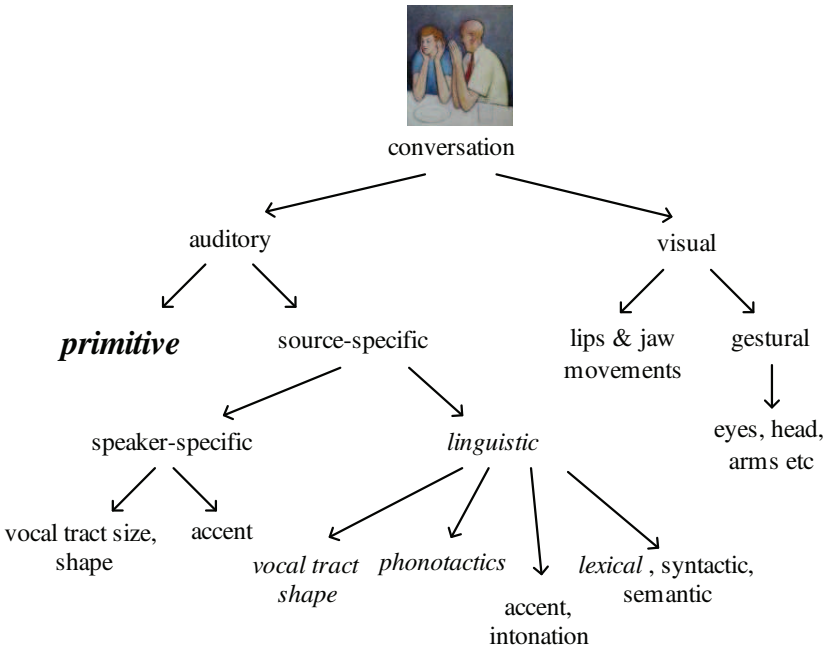


Fig. 2. Potential cues that are useful for separation. This picture is adopted and redrawn from Martin Cooke's NIPS 2006 tutorial presentation.

The differences in source characteristics are exploited, rather than individual characteristics. Consequently, speech separation is suitable for dynamic environments of sources with rapid-changing characteristics.

Computational auditory scene analysis (CASA) is one of the popular speech separation methods that exploits human perceptual processing in computational systems (Wang & Brown, 2006). Human beings have shown great success in speech separation using our inborn capability. Our perceptual organization always gives reliable performance that individual sound sources are resolved, even with a single ear (Bregman, 1990; Cherry, 1953). The separation remains effective, even for sound sources with fast-changing properties, such that appropriate actions can be taken by knowing the present environment around us. All of these suggest that modeling how human being separates mixed sound sources is a possible way for speech separation.

Given an input mixture observation, it is undergone an auditory scene analysis (ASA), which first examines various cues from the mixture observation (Bregman, 1990). Cues are related to the rules that govern how sound components from one source should look like. Take an example, voiced speech source has power always located at multiples of the fundamental frequency (F0). These frequency components are grouped together by using the harmonicity cue (Bregman, 1990). Furthermore, cues are associated with some relevant features. F0 and power distribution are the features for the cue mentioned above. After applying different cues, sound components, which are likely to come from the same origin, are grouped together as one single source. These resultant sources finally constitute the 'scenes' that experienced.

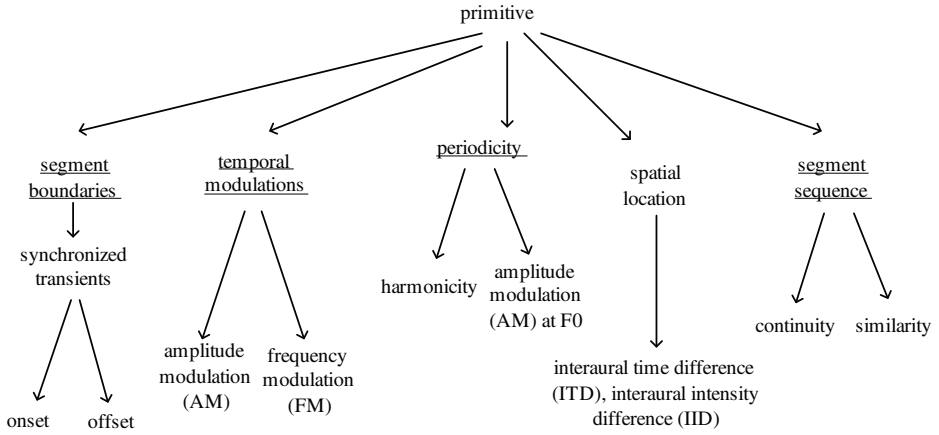


Fig. 3. Primitive cues (This picture is adopted and redrawn from Martin Cooke's NIPS 2006 tutorial presentation). For single-microphone speech separation, cues underlined are available.

A number of cues are applied during ASA. They are either primitive or schema-based. Fig. 2 and 3 depict some potential cues for separation. Primitive cues are global, physical constraints, such as those structural properties of sound sources, source-listener geometry, etc., which lead to certain acoustic consequence. Harmonicity and continuity are two examples. They are independent of the speech source identity and remain valid in a wide variety of auditory scenarios. Schema-based cues are source-specific. They are related to speakers' characteristics and the linguistic aspects of the languages. Examples for schema-based cues are context information, prior knowledge of familiar auditory patterns, etc. (Bregman, 1990). Those in italics (Fig. 2) are cues that adopted in the separation algorithm introduced later.

Regarding linguistic cues, different aspects, namely vocal tract shape, accent, intonation, phonotactics, lexical, syntactic and semantic rules are included. By phonotactics, it refers to a branch of phonology dealing with restrictions in a language on the allowed combinations of phonemes. Lexicon is the vocabulary of a language. Syntactic rules govern the ways of sentences are formed by the combinations of lexical items into phrases. Semantic rules concern about meanings are properly expressed in a given language.

Traditionally, primitive cues are often adopted in separation methods (Brown & Cooke, 1994; Hu & Wang, 2004), due to the use of simple tonal stimuli in perceptual studies and their reliability over time.

The first separation method appeared in 80's. Parsons proposed a frequency-domain approach for extracting target voiced speech sources from a mixture of two competing voices (Parsons, 1976). Strong influence of harmonicity cue on auditory perception (Assmann, 1996; Meddis & Hewitt, 1992) are observed in voiced speech. For two voiced speech sources, the mixture observation is essentially the sum of the comb-like harmonic spectra. This separation process is aimed to compile a set of spectral peak tables which describe the frequency, magnitude, phase and the origin of each spectral peak and reconstruct individual voiced sources accordingly. Overlapped peaks are resolved and source F0s are tracked. As harmonicity is the only cue involved and this method is limited to vowels and fully voiced

sentences, separation methods later on move to apply more and more cues on various types of speech, which are present in real conversation.
Weintraub later proposed a CASA system for the separation of simultaneous speech of a male and female speaker (Weintraub, 1985). Harmonicity and continuity are employed. Continuity refers to the auditory behaviour that a continuous trajectory or a discontinuous but smooth trajectory tend to promote the perceptual integration of the changing experience (Bregman, 1990). The change could be in terms of frequency components, pitch contour, intensity, spatial location, etc. On the other hand, an abrupt change indicates that a new source appears. This system consists of three stages:

1. The pitch periods of individual speakers are first determined by dynamic programming (DP). Output power of each frequency channel is inspected and summaized. The pitch period with the strongest peak is selected and assigned to the speaker who has the closest average pitch (Data with annotations are collected from four speakers).

   The resultant pitch periods are potential pitch values only. The next stage determines if the individual sources are voiced or unvoiced at different moments.

2. The number of periodic sources and the associated type are decided by a pair of Markov models by using the continuity cue. One Markov model is dedicated for one speech source, which consists of seven states: silence, periodic, non-periodic, onset, offset, increasing periodicity and decreasing periodicity. The Viterbi algorithm (Jelinek, 1997; Viterbi, 2006) is used to find out when a sound source starts to be periodic and when it becomes non-periodic.

3. After knowing the number and the characteristics of sound sources over time, the magnitude in each channel for a given sound source is estimated. By looking at pre-computed records which store the expected magnitude in each channel with the pitch frequency information, individual magnitudes are estimated.

Shortly after the publication of the CASA tome from Bregman, a series of separation systems for speech sources or music emerged (Brown, 1992; Cooke, 1993; Ellis, 1994; Mellinger, 1991). All of them differed from previous systems in a way that a number of primitive grouping cues are adopted, rather than exploiting harmonicity alone. Table 1 lists out the primitive cues used. These systems basically follow the framework below. After time-frequency analysis and feature extraction, sound components with coherent properties are searched and grouped together, according to the results from a variety of cues. These systems are 'data-driven' such that the separation result solely depends on the primitive cues of the mixture signal. No prior knowledge or learned pattern is involved.

system	harmonicity	continuity	onset	offset	AM	FM
Weintraub (Weintraub, 1985)	✓	✓				
Cooke (Cooke, 1993)	✓	✓			✓	
Mellinger (Mellinger, 1991)		✓	✓			✓
Brown (Brown, 1992)	✓	✓	✓	✓		

Table 1. The use of various primitive cues in different CASA systems.

The study of Brown presents a segregation system that improves from previous systems (Brown, 1992; Brown & Cooke, 1994; 1992). It is aimed to segregate speech from an arbitrary intrusion, such as narrow-band noise, siren and concurrent speech source. Primitive features

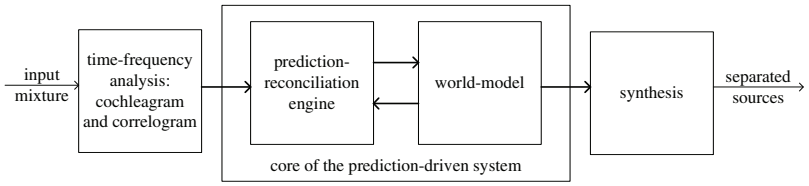


Fig. 4. Block-diagram of Ellis' system, showing the prediction-reconciliation engine looks for a prediction constructed by elements inside the world-model, such that the observations in energy envelope and periodicity are matched.

including harmonicity, continuity, onset and offset are computed from the auditory nerve activity. Segments are derived by identifying contiguous channels having similar local harmonicity. They are then aggregated over time by finding the most probably pitch contour using dynamic programming (Cooper & Cooper, 1981). For segments with synchronous onset or offset, they are likely to be assigned to the same source.

Several years later, a 'prediction-driven' separation system is proposed by Ellis (Ellis, 1996). This is a schema-based system. It comprises a process of reconciliation between the observed low-level acoustic features and the predictions of a world-model. The world-model constructs a simplified representation of an input stimulus (acts as an external world) by using a collection of independent sound elements.

The collection consists of three elements: noise clouds, tonal elements and transients. They are chosen as 'schemas', so as to reasonably model arbitrary signals and give satisfactory expression of subjective experience. The noise cloud class is used to represent 'noise-like' sounds with specific intensity over a range of time and frequency, but without a fine structure. Tonal elements simulate sounds that give perceptible periodicity. They have a pitch track besides the time-frequency energy envelope that characterizes a noise cloud. The third class of sound element is designed to model typical clicks and cracks. These rapid bursts of energy differ from the previous two classes in a way that no periodicity is perceived, being short in duration with a broad spectrum and followed by a fast exponential decay of energy.

If the predicted sound events is found to be consistent with the extracted cues, they will form a scene interpretation, no matter if direct primitive evidence is observed or not. Fig. 4 depicts the basic layout of Ellis' prediction-driven system.

Recently, Barker *et al.* has proposed a recognition algorithm (Barker et al., 2005) which decodes speech signals from a segregation perspective. It is aiming at finding the word sequence, along with the time-frequency segregation mask for a given noisy speech observation, that is,

$$\hat{W}, \hat{S} = \arg\max_{W,S} P(W, S|\mathbf{Y}) \tag{1}$$

where W and S are the word sequence and segregation mask respectively. $\mathbf{Y}$ are the observation features. By including an unknown speech source feature term $\mathbf{X}$, they expressed $P(W, S|\mathbf{Y})$ in terms of a hypothetical segregation term $P(S|\mathbf{Y})$. Barker *et al.* evaluated this extraction algorithm on a connected digit task. The observation features $\mathbf{Y}$ are from corrupted speech with background noise. Comparing with a conventional recognition system, a significant reduction in word error rate is achieved.

Schema-based separation is an emergent and potential direction that research work on separation algorithm is less explored, compared to systems using primitive cues alone (Barker

et al., 2005; 2006; Brown, 1992; Brown & Cooke, 1994; Cooke, 1993; Ellis, 1996; Hu & Wang, 2006). Traditional CASA separation systems rely on the use of primitive cues in a data-driven manner, where low-level acoustic features are gradually merged to form the resultant scenes. Nevertheless, schema-based separation is indispensable and advantageous to the underdetermined nature of single-microphone speech separation. Perceptual experiments have shown that using primitive, acoustic features alone in a bottom-up architecture is simply insufficient to achieve the superior performance of the human auditory system that we habitually possess. Schema-based separation with speech modeling is expected to be beneficial and offer linguistic cues in a top-down direction.

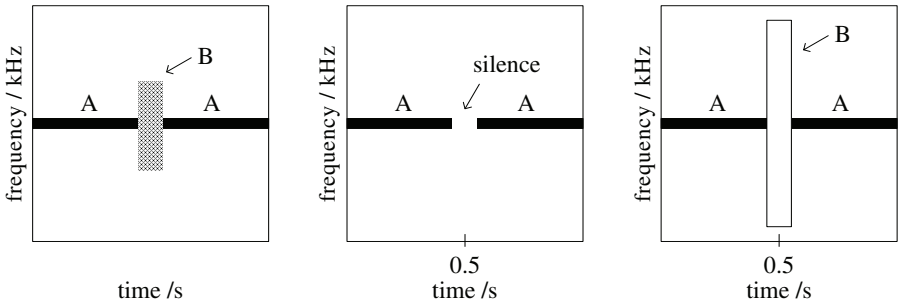
Speech signals are much complex than tonal signals. Moreover, the psychoacoustic studies are generally not aimed at designing separation methods. Are these human perceptual principles applicable in automatic speech separation as well? What are their relative strengths in deriving the 'scenes'? What are the appropriate representations for these cues to be incorporated in separation methods? In the following, perceptual cues and the associated psychoacoustic studies are analyzed from the perspective of speech separation. Potential cues are compared and selected based on their merits for separation.

2. Perceptual cues for speech separation

2.1 Harmonicity

Harmonicity refers to the property that if a set of acoustic components are harmonically related, i.e. having frequencies that are multiples of a F0, listeners tend to group these components into a single scene. This property is sometimes referred as periodicity. This demonstrates that the pitch frequency of a speech source is one of the essential factors influencing our separation. A number of work have attempted to study this on perceptual segregation of double (simultaneous) synthetic vowels (Assmann, 1996; Assmann & Paschall, 1998; Assmann & Summerfield, 1990; Meddis & Hewitt, 1992; Scheffers, 1983). In one of Scheffers's experiments, five synthetic vowels with 200 ms duration were involved. It is found that when the vowels were unvoiced, or both with the same F0, listeners were able to identify both constituents significantly better than chance level. Furthermore, when a difference in F0 was introduced, the identification rate of both vowels improved. Identification performance increased with the increase of F0 difference and reached an asymptote at 1-2 semitones. This finding is consistent to many other research work, showing that listeners are able to identify both vowel pairs with a performance is significantly above chance, even when the two vowels are with similar amplitude, starting and ending time and presented to the same ear. From this experiment, the harmonic patterns of periodic sounds is expected to provide proper separation of concurrent sounds. This is particularly useful, because speakers are unlikely to share identical F0 at the same instant.

The above Scheffers' experiments have clearly shown that harmonicity determines the grouping of sound components, as long as the constituent sources are voiced and with distinct F0s. Nevertheless, the identification remains effective and significantly better than chance level, even when the vowels are unvoiced or share the same F0. For this case, harmonicity is inapplicable. This reflects that some other cues are involved, besides harmonicity, where speech models developed over time are likely to contribute.



(a) a softer signal A with an intense interrupt B

(b) tone filled with alternating silence

(c) tone filled with alternating noise burst

Fig. 5. An perceptual experiment about the continuity cue, redrawn from (Bregman & Ahad, 1996).

2.2 Continuity

Recall that continuity refers to the auditory behaviour that a continuous trajectory or a discontinuous but smooth trajectory tend to promote the perceptual integration of the changing experience (Bregman, 1990). On the other hand, an abrupt change indicates that a new source appears. The continuity cue is based on the Gestalt principles (Bregman, 1990; Koffka, 1963). Dates back to the last early century, a group of German psychologists formulated a theory of perceptual organization. This theory is aimed to explain why patterns of highly connected visual objects are formed, despite sensory elements seem to be separated. The German word 'Gestalt' means pattern. The Gestalt principles apply not only to images, but also to sounds.

This continuity cue has been investigated by many perceptual experiments (Bregman, 1990; Bregman & Ahad, 1996; Bregman & Dannenbring, 1973; Ciocca & Bregman, 1987; Darwin & Bethell-Fox, 1977). Different types of sound materials were used, including synthetic vowels, alternating tone sequences, tones preceding and following a noise burst and so on. These experiments have demonstrated the effect of continuity on the unitary perception of sound components, regardless of simple tones or complex speech sources.

To explain the principles how continuity is used during ASA, Fig. 5 shows a classical perceptual experiment studying the perceptual continuation of tone through a noise burst (Bregman & Ahad, 1996). Part of a signal A is deleted and replaced by a short, louder interrupting sound B. A simple example would be the following: Signal A is a 1000 Hz pure tone and the interrupting sound B is a burst of white noise. The tone is presented with alternating silence at the beginning (Fig. 5(b)), which consists of 900 ms of tone with 100 ms of silence in one cycle. After several cycles, the noise is introduced and after each group of cycles, the noise intensity is increased. In the final group of cycles (Fig. 5(c)), the spectrum level of the noise is the same as tone A at 1000 Hz. Listeners now hear the tone as continuously unbroken behind the noise burst. Furthermore, this does not happen in any of the previous intensities.

This perception of continuity occurs when the neural activity during B includes activity very similar to what normally occur in signal A. Hence, there is evidence that signal A may still be present behind the interrupt. If this evidence is missing, for example, if there is a sudden change at the boundary, then listeners would perceive signal A as a signal that only lasts until the interruption. In the experiment above, continuity comes when the noise level is the same

as the tone level in the final group of cycles. During listening, the auditory system tries to detect if the signal A continue inside B or if there is any sudden change that destroys the continuation. In a more complex case, a spoken sentence appears to continue through a short loud noise burst (Bregman, 1990; Warren, 1970; 1982; Warren & Obusek, 1971; Warren et al., 1972). This 'phonemic restoration' is due to the fact that our ASA process often restores the perception of a phoneme, so as to complete the partially deleted word.

The continuity cue has not been widely adopted in current CASA systems. There are only a few examples using it as a supplement to other popular cues, such as harmonicity, onset and offset (Cooke & Brown, 1993; Klapuri, 2001). A possible reason for this is given below, which is related to the natures of signal A in the above perceptual experiment. The continuity principle monitors the degree of restoration of signal A inside the occlusion. This leads to a question: Is it necessary to locate signal A before applying the continuity cue to have restoration? According to the continuity condition described in the perceptual experiment, ASA determines if there is evidence that the on-going signal A presents inside the interrupt. Hence, the boundaries, that is, the occlusion part, and the parts before and after it must be defined first. This becomes a prerequisite and the continuity cue acts as a supplementary cue to predict the missing part only. Since the boundaries of occlusion is already defined, other top-down information, for example, lexicons, phonotactics and syntactic constraints can be derived accordingly and used to replace the continuity cue. Compared with the continuity cue, these source-specific cues are global and statistical modeling is ready made. Furthermore, to define whether a change is with good continuation is not an easy task by itself, in particular, for speech signals with varying properties over time. When compared with other grouping cues, such as harmonicity, the description of continuity is relatively loose. Unlike continuity, constitutive components of a harmonic source share a strong mutual relation which are clearly defined in physical laws. Given a signal A, it remains unknown to decide if the trajectory has a close cubic polynomial fit is continuous or not. It is a relative issue that considers physiological limitations and natures of signal A.

2.3 Onset & offset

This onset & offset cue is under the Gestalt principle of common fate (Bregman, 1990; Koffka, 1963). This principle states that our perceptual system tends to group sensory elements which synchronously change in the same way, acting as a way of organizing simultaneous sounds. In reality, frequency components originated from the same source often vary in a coherent manner. For instance, they tends to start and stop together (sharing an identical time record), change in amplitude or frequency together (results in amplitude or frequency modulation). On the other hand, components from independent sound sources rarely start or end at the same time in normal listening environments. Onset and offset correspond to the beginning and ending time of a sound source respectively. They are especially influential to voiced speech and music (Mellinger, 1991).

Psychoacoustic evidence of the onset and offset principle has been intensively investigated in tone, speech and music stimuli (Bregman & Ahad, 1996; Bregman & Pinker, 1978; Dannenbring & Bregman, 1978; Darwin, 1984; Darwin & Ciocca, 1992; Darwin & Sutherland, 1984; Summerfield & Culling, 1992). In one of the Bregman's experiments (Bregman & Pinker, 1978), a paradigm as shown in Fig. 6 is adopted to illustrate the effect of onset and offset synchrony on streaming.

The stimulus used for the experiment is a repeating cycle composed of a pure tone A alternating with a complex tone with components B and C. A, B and C are pure tones with

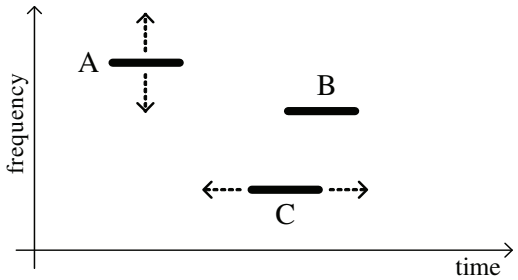


Fig. 6. A demonstration showing the influence of onset and offset synchrony on streaming (redrawn from (Bregman & Pinker, 1978)). Frequency proximity and synchrony of tones determine how the three tones are perceived.

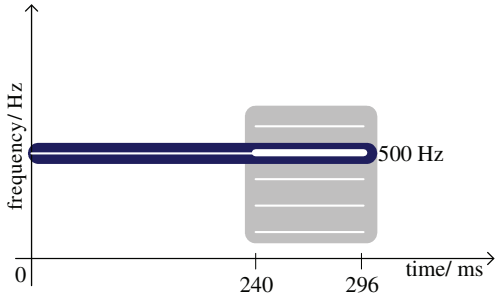
identical amplitude and equal duration. The frequencies of B and C remain unchanged throughout the experiment. From the experimental results, it is shown that simultaneous and sequential groupings compete with each other. Two factors are involved in the competition, namely the frequency proximity of A and B and synchrony of B and C. If B and C share common onset and offset, this synchrony strengthens the force of simultaneous grouping of B and C to become a complex tone (BC-BC-BC-BC), which would otherwise belong to separated streams as in the case of sequential grouping. In this experiment, both onset and offset of B are made to be synchronous or asynchronous with those of C at the same time. If only the offset synchrony is employed alone, its influence is found to be much weaker than the onset synchrony (Brown, 1992; Darwin, 1984; Darwin & Sutherland, 1984).

Although experiments have consistently reported that onset and offset are influential to perceptual grouping, they are not sufficient conditions for grouping the harmonics of a vowel (Barker, 1998; Darwin & Sutherland, 1984). They are not necessary conditions neither. In some cases, the onset and offset cues are not necessary, for example, the experiment described under the harmonicity cue. Concerning whether they are sufficient conditions or not, further experiments on voiced speech illustrate the arguments. Darwin and Sutherland have demonstrated that leading onset of a harmonic can reduce its contribution to the vowel's quality (Darwin & Sutherland, 1984) (as shown in Fig. 7(a)). However, if a tone at the octave of the harmonic which onsets synchronously with the leading harmonic, but stops as the vowel begins is added, the reduced contribution of the leading onset will be canceled. The extent to which an individual harmonic contributes towards the perception of a vowel depends on various factors: (1) whether it starts at the same time as other harmonics; (2) whether it ends at the same time; (3) whether it forms a perceptual group with other sounds. Referring to Fig. 7(b), the additional octave tone and the leading section of the harmonic jointly form a separate stream that ends just before the vowels begins. The remaining section of the harmonic is left to be part of the vowel perception.

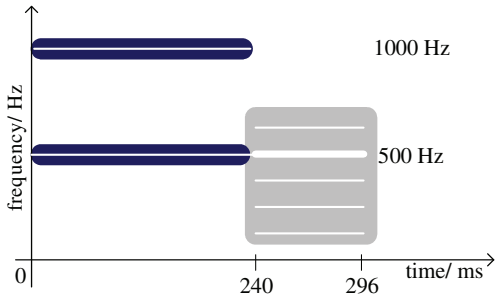
Hence, it is believed that other principles and grouping cues are expected to be involved. Take an example, harmonicity acts as a stronger cue and onset and offset are comparatively supplementary in this experiment.

2.4 Amplitude modulation & frequency modulation

The cue that have just been discussed — onset & offset accounts for instantaneous events which happen shortly. A more general Gestalt principle of common fate concerning the



(a) A vowel with asynchronous harmonic onset



(b) A vowel with asynchronous harmonic onset, together with a octave tone

Fig. 7. Contribution of a leading onset of a harmonic to the vowel perception(redrawn from (Darwin & Sutherland, 1984)). Different streams are highlighted with distinct colors.

change over the course of time is the amplitude and frequency modulation (Bregman, 1990; Darwin, 2001; Koffka, 1963). This cue refers to the tendency that if any subset of elements of an auditory scene changes proportionally and synchronously, it will be segregated from other elements within the scene that are changing in a different way. It is very unlikely to have unrelated elements within an auditory scene undergoing parallel changes by accident. It is much more reasonable to assume that these unrelated elements are arisen from the same source. For a continuous speech source, frequency components will go on and off roughly at the same time, glide up and down in frequency together (called frequency modulation or FM), be amplified and attenuated together (called amplitude modulation or AM) and so on. These latter two types of synchronous changes have been studied in various psychoacoustic experiments with complex tones or vowels (Bregman, 1990; Bregman et al., 1985; Bregman & Ahad, 1996; Bregman et al., 1990; Carlyon, 1991; Darwin, 2001; Gardner et al., 1989; Hall et al., 1984; Kubovy, 1981; Marin & McAdams, 1991; McAdams, 1989; Purwins et al., 2000; Summerfield & Culling, 1992).

There are experiments showing evidence on the use of common AM for grouping, which requires frequency elements to be fused sharing identical modulation frequency and phase. The effects being observed, however, are small (Bregman et al., 1985). Other experiments on amplitude modulation show equivocal evidence. Listeners are found to be unable to take

advantage of AM incoherence for slow modulating rates like 2.5 Hz (Darwin & Carlyon, 1995; Summerfield & Culling, 1992).

Concerning frequency modulation, human voice is one of the everyday examples. It is found that the components of all natural sounds with sustained vibration contain small-bandwidth random fluctuations in frequency. Speech voices and sounds from musical instruments share this property (Deutsch, 1999; Flanagan, 1972; Jansson, 1978; Lieberman, 1961; McAdams, 1989). For real speech signals, the pitch frequency varies from time to time. When the vocal cords are tightened, the pitch frequency increases. Simultaneously, all the harmonics rise in a proportional way, so as to maintain their constant ratios to the fundamental. Furthermore, some experiments have shown that both harmonic and inharmonic stimuli are perceived as more fused when coherent modulation is applied (Helmholtz, 1954). Due to this coherence among the frequency components, it is possible that the auditory system employs such coherence frequency modulation as a cue for grouping spectral components together; and conversely treat any spectral components with different frequency modulation as separate streams.

To test this hypothesis, McAdams designed a series of experiments which studied the effect of frequency modulation on segregation of concurrent speech sources (McAdams, 1989). It is found that when frequency modulation is present on a vowel, the perceived prominence of the vowel increases; however, this modulation effect is independent of the modulation states of any accompanying sounds. The prominence is not reduced when accompanying sounds share identical modulation with the vowel. There is no change in vowel prominence if accompanying sounds are modulated incoherently. The last condition is highly similar to the scenario with multiple speech sources. Frequency components of a continuous speech source rise and fall over time and components from different sources are modulated incoherently. Although modulation to the target source aids to promote its prominence, there is no distinction between coherent and incoherent modulation to humans. Therefore, common FM does not help segregation of the components from a speech source from a mixture signal. Subsequent experiments further support the absence of effects of common FM or AM during ASA (Carlyon, 1991; 1992; Deutsch, 1999).

Stimuli consists of three simultaneous vowels (/a/, /i/ and /o/) at different pitches in a chord of two seconds duration. Four conditions of coherent or incoherent frequency modulation have been tested: (1) no vowels are modulated; (2) all vowels are coherently modulated by an identical modulating signal; (3) one vowel is modulated, while the other two are remained steady without any modulation; (4) one vowel is modulated independently against a background of the remaining two vowels, which are modulated coherently with one another. The modulating signal is constituted by both periodic and noise-like components. It has been verified in a pretest that these individual vowels are identifiable in isolation with perfect accuracy, regardless of pitch and the presence or absence of modulation.

Another explanation for the absence of contribution of common FM to segregation is based on the close relationship of FM with harmonicity. Continuous voiced speech with rising or falling F0 exhibits not only common FM, but also the harmonicity over time. Whether the coherent motions of frequency components or their harmonic structure contribute, indicates if they can be used as basis for segregation. Summerfield and Culling has investigated the role of harmonicity and common FM by performing experiments using different F0 and FM (Summerfield & Culling, 1992). They also checked if the harmonicity cue is so strong that most of the segregation is already achieved, leaving nothing for other cues to contribute. Based

on the experimental findings, they concluded that common FM cannot help segregation of concurrent sources and harmonicity cue is always dominant.

2.5 Schema-based grouping

The following addresses the use of high-level cues for separating monaural speech. Specifically, the use of linguistic and visual information in speech perception will be discussed. Primitive cues arise from acoustic properties of the environment, however, there exists many observations that are contributed by either attention or learning (Bregman, 1990; Dowling, 1973; Meddis & Hewitt, 1992; Scheffers, 1983; Warren, 1970). To account for the observed human capabilities in these experiments, primitive cuess are inadequate. Dowling showed that the identification of interleaved pairs of melodies is easier when the melodies are familiar or when music training is acquired beforehand (Dowling, 1973). Attention(conscious effort) can also be used to specifically hear out frequency components from complexes that would group together otherwise. Based on these experimental findings, it is evident that some higher-level information other than primitive features is involved.

The double synthetic vowel experiment described under the harmonicity cue also demonstrates that primitive cues alone cannot justify the phenomenon that listeners are able to identify the constituent vowels significantly above chance, even when nearly all primitive features are approximately the same, identical amplitude, onset and offset time and F0. The difference between the two constituent vowels lies only in their spectral shapes. This demonstrates that even when the grouping mechanism is unable to separate the incoming sound using primitive cues, the innate recognition processing is still responsible for recognizing each of the two overlapping constituent vowels and generate separated 'scenes'.

Listeners are capable of segregating sound sources, in particular, speech sources under situations where certain primitive features are difficult or impossible to be observed from the input mixture signal. During the perceptual organization, by guessing the missing or occluded information with some high-level information, for example, those linguistic cues that listed in Fig. 3, the estimate is instinctively incorporated as if it has been directly perceived. In the following, evidence of the schema-based grouping from experimental psychoacoustics will be reviewed, which illustrates what logical basis have been adopted.

One of the examples is about this restoration phenomena from Warren (Warren, 1970; 1982; Warren & Obusek, 1971; Warren et al., 1972; Weintraub, 1985). Listeners hear a speech recording. Within the utterance, an entire phoneme has been removed and replaced by a loud noise burst, which sounds like a cough. 'Auditory induction' then occurs, where the missing phoneme is perceived on top of the loud noise. Based on the contextual cues of neighboring regions (linguistic constraints), listeners unconsciously infer the most likely sound. Furthermore, the precise location of the missing part could not be identified.

Subsequent studies have shown that the success of phonemic restoration is contributed by several factors, including repeated listening of the speech utterance, prior knowledge and linguistic skills of listeners, timbres of different sound sources, the characteristics of the noise masker and the associated continuity.

Another way in which the auditory system uses linguistic information, particularly in the application of speech separation, can be seen in the following example. Cherry has performed a set of recognition experiments using speech stimuli, to investigate the ways that we recognize what one person is saying while the others are talking simultaneously (Cherry, 1953). Specifically, the objective is to study if linguistic knowledge promotes speech separation. To focus on linguistic cues, other factors, such as, lip-reading, gestures, spatial

geometry and speaker-specific styles are eliminated. This is done by adding pieces of speech recording spoken by the same speaker together to form the stimuli. As a result, only linguistic cues, for example, transitional probabilities between phonemes (phonotactics), subject matter (semantics and lexicon), syntax and intonation and monaural primitive grouping cues remain. The experiments are designed to present two mixed speech passages to a listener and he is asked to separate one of the passages and verbally repeat it word by word or phrase by phrase. Most of the listeners' extracted speech is found to be accurate, with little errors only. There are few transpositions of phrases between passages which are highly probable. In some passage pairs, there is no such transposition exists. Besides, there is no phrase with consecutive two or three words being wrongly recognized.

It is believed that this recognition task requires the knowledge of neighboring events (the events can be phonemes, words and subject matters) and a big storage of probabilities, calculating the probability rankings of various phonemic and lexical combinations. While the knowledge of neighboring events could be resulted from primitive grouping or previous learning, the storage of probabilities enables prediction made on a probabilistic basis, so as to combat any interference or occlusion from competing speech.

Schema-based grouping is powerful, particularly when some critical acoustic features are made to be unobservable by interfering signals. Hypothesis from high-level knowledge compensates any missing components and removes redundant elements according to the relevant continuously-trained models. Comparatively, primitive grouping does not have such compensation or removal functionality, since missing components are simply absent in the mixture input and primitive grouping just assigns (part of or the whole) sound components to either source. Nothing will be added or taken away during grouping.

Nevertheless, schema-based organization is not independent of primitive grouping. They rather collaborate in most sound separation scenarios. Schemas are developed by learning common patterns. At the early stage, primitive grouping provides initial organization, which is then left for the interpretation and 'bootstrap' of the schemas. When familiar patterns are gradually developed, features from primitive grouping cues enable reduction in the hypothesis space that handled in the schema-based organization. After the optimal hypothesis is determined, grouping at primitive level is changed accordingly and this process may be repeated, until a consistent grouping is achieved.

Knowing that harmonicity and prior knowledge of familiar auditory patterns for schema-based grouping are found to be powerful, whereas other cues such as continuity remain relatively inferior. In the following section, a separation algorithm using the idea of recognition models is introduced. It incorporates the statistical distribution of speech sounds in a top-down direction to perform separation and predict target source signals in terms of low-level features. Models of individual speech sounds are employed to supply any missing components and eliminate redundant ones.

3. Speech separation with speech models

A speech separation algorithm for single-microphone mixture input is introduced (Lee et al., 2007). It showcases a potential way to utilize both primitive and schema-based cues in separation systems. The meaning of 'model' is two-folded. First, based on the source-filter model of speech signals, the algorithm reconstructs individual speech sources by estimating their associated spectral envelopes and suppress the interfering source accordingly. Second, with the use of speech models as prior knowledge, trajectories of spectral envelope are

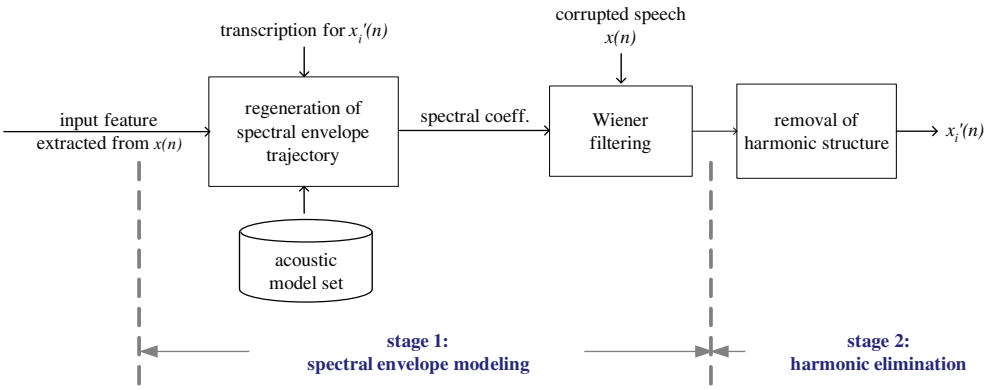


Fig. 8. Block-diagram of the proposed separation system. Same procedure is applied to all speech sources.

regenerated in a top-down, probabilistic manner. Selective primitive cues are incorporated as well to generate smooth source outputs with single periodicity (Lee et al., 2006).
Speech modeling and recognition techniques are adopted to statistically capture this knowledge and 'familiar auditory patterns' are built. They are used to govern the separation process (in a top-down, macroscopic view) and to revise typical primitive features of source estimates (in a bottom-up, microscopic view). These statistics of auditory patterns acts as prior knowledge to the proposed separation system and are gradually updated during the training process for the embedded speech recognition engine of the proposed system.
Fig. 8 depicts the block-diagram of the proposed separation system, where $x_i'(n)$ is the estimate for source i ($i \in [1, 2]$). To extract either of the speech sources, $x_1(n)$ or $x_2(n)$, same procedure is applied with slightly different inputs. It is assumed that the text transcription for both speakers are either known or estimated by some means.

1. The first stage is about extracting the spectral envelope of target source. A speech recognition engine is employed inside the block 'regeneration of spectral envelope trajectory', so as to output the optimal model sequence. At the output end, a sequence of spectral coefficients describing the expected spectral envelope trajectory for the target source $x_i(n)$ is estimated.
2. Wiener filtering is then used to remove the interference source, based on the spectral information of the target source obtained from the preceding regeneration block. The resultant spectral envelope behaves properly with peaks at target formant frequencies. This completes the envelope extraction part.
3. Provided that both $x_1(n)$ and $x_2(n)$ are voiced in the current frame, the excitation sources of both target and interference are still present at the Wiener filter output. In the second stage, a comb filtering block is used to remove any harmonic structure associated with the interfering source. This completes the harmonic part and the output waveform will be the estimated source $x_i'(n)$.

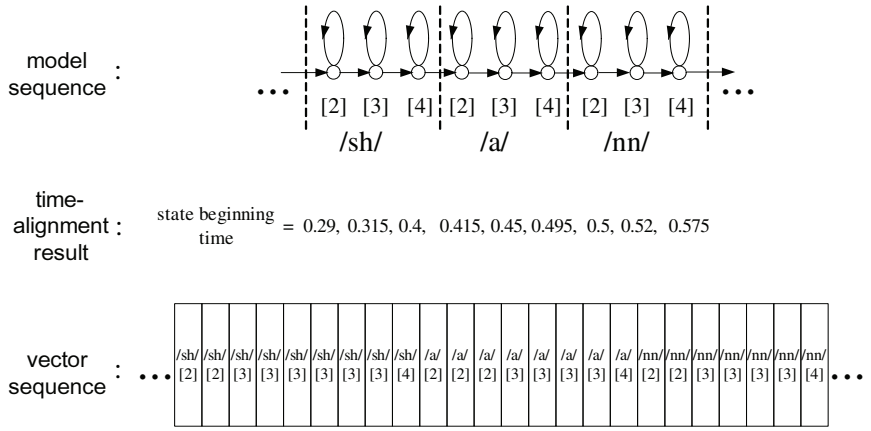


Fig. 9. An example of a time-alignment result and the corresponding vector sequence. The symbols /·/ and [·] denote model name and state number respectively. Each HMM contains five states, with state [2], [3] and [4] are emitting states. The longer a state is aligned, the more the mean vector is repeated.

3.1 Regeneration of spectral envelope trajectory

To search for or align with any familiar patterns of the input mixture signal $x(n)$, speech models in the form of reliable feature representation are necessary. Line spectrum pair (LSP) ((Itakura, 1975; Paliwal, 1989; Soong & Juang, 1984; Wakita, 1981)) is employed to parameterize $x(n)$ and a set of hidden Markov models (HMMs) in phone units is trained to model these familiar patterns.

The input mixture signal $x(n)$ is first converted to a sequence of LSP feature vectors. With a given transcription, the 'regeneration of spectral envelope trajectory' block uses the Viterbi algorithm (Jelinek, 1997; Rabiner & Juang, 1993) to perform forced alignment at state level. By storing a set of familiar speech patterns (acoustic models), this regeneration block helps to retrieve the expected spectral shape of a target speech source, $X_i(\omega, t)$. Moreover, the time-alignment defines the boundaries when a particular model state is activated. The mean vectors of associated acoustic models are extracted from the model set and replicated according to the time-alignment. By looking at this vector sequence, the expected spectral shape evolution of the target source, $X_i(\omega, t)$ is illustrated in LSP sense.

Fig. 9 shows an example of a time-alignment result and the respective vector sequence. Suppose acoustic models /sh/, /a/ and /nn/ are retrieved and aligned with the observed sequence of LSP feature vectors. After alignment or recognition, the activation time (time-alignment) of individual model states is output in the form of the beginning time and the ending time. In the current implementation, the state beginning time is used only, but not the state ending time. This is because the beginning time of a model state is just the next time unit of the ending time of the preceding model state. As alternative, the ending time record can be used instead. Referring to the time-alignment result (shown in the middle), state [2] of /sh/ is activated at 0.29 s and state [3] of /a/ is switched on at 0.45 s. Consequently, by putting the speech-pattern describing model mean vectors together accordingly, the longer a state is aligned, the more the model mean vector is repeated, the spectral trajectory is approximated.

The resultant vector sequence will be similar to the one shown at the bottom, which represents the expected spectral shape evolution of the target source, $X_i(\omega, t)$.
The generated trajectories are piecewise constant, since model mean vectors are just kept intact and concatenated together. This, however, introduces (1) discontinuities at state transitions; and (2) conflicts between the static and dynamic coefficients. Hence, based on the statistics of dynamic coefficients modeled by HMMs, individual trajectories of static components are further revised before leaving the 'regeneration of spectral envelope trajectory' block in Fig. 8 and smooth, consistent trajectories are generated. This revision is carried out in the LSP domain, since the LSP representation for ordered and bounded coefficients is robust to interpolation and quantization. Finally, the resultant coefficients representing the expected spectral envelope of a target source are converted back to LPC coefficients $\{a_m\}$ as the output.
Familiar speech models, in the form of mean and variance vectors of static and dynamic LSP coefficients, are trained by a set of speech training data. The model set is HMM-based.
The static component of the vector sequence obtained from Viterbi exhibits the spectral shape evolution of a target source. It is also associated to the dynamic component by linear regression. Simple replication of the mean vector from a model state will produce piecewise constant static components in consecutive frames, but non-zero dynamic components. Nevertheless, this inconsistency does not exist in real speech signals. As a result, the dynamic component, including the delta and acceleration coefficients, is used as a constraint in generating the static component, similar to the ways showed in (Lee et al., 2006).
Let $\theta_{im}(n), m = 1, 2, \ldots, M$ be the set of LSP coefficients estimated at frame n for source i. The order of LSP is denoted by M. By putting these θ_{im} vectors together, the spectral shape evolution for source i is exhibited. Consider the m-th dimension of the vector sequence, the temporal trajectory is expressed as

$$\boldsymbol{\theta_{im}} = \left[\theta_{im}(1)\ \theta_{im}(2)\ \cdots\ \theta_{im}(N)\right]^T \tag{2}$$

where N is the number of frames in total. Similarly, we have

$$\Delta\boldsymbol{\theta_{im}} = \left[\Delta\theta_{im}(1)\ \Delta\theta_{im}(2)\ \cdots\ \Delta\theta_{im}(N)\right]^T \text{ and} \tag{3}$$

$$\Delta^2\boldsymbol{\theta_{im}} = \left[\Delta^2\theta_{im}(1)\ \Delta^2\theta_{im}(2)\ \cdots\ \Delta^2\theta_{im}(N)\right]^T \tag{4}$$

where $\Delta\theta_{im}(n)$ and $\Delta^2\theta_{im}(n)$, for $n = 1, 2, \ldots, N$ are the delta coefficients and acceleration coefficients respectively. T represents the transpose operation.
For each LSP dimension, the contour $\boldsymbol{\theta_{im}}$ is generated by finding the maximum-likelihood estimate by the aligned static, delta and acceleration statistics. Equivalently, this is given in Equation (5).
The regenerated trajectory is found by,

$$\boldsymbol{\theta'_{im}} = \left[\mathbf{W_{LSP}}^T \mathbf{\Sigma_{im}}^{-1} \mathbf{W_{LSP}}\right]^{-1} \mathbf{W_{LSP}}^T \mathbf{\Sigma_{im}}^{-1} \boldsymbol{\theta_{im}} \tag{6}$$

where $\mathbf{\Sigma_{im}}$ and $\mathbf{W_{LSP}}$ are the diagonal covariance matrix and

$$\mathbf{W_{LSP}} = \begin{bmatrix} \mathbf{I} \\ \mathbf{W} \\ \mathbf{W^2} \end{bmatrix} \tag{7}$$

$$\boldsymbol{\theta_{im}} = \begin{bmatrix} \boldsymbol{\theta_{im}} \\ \Delta\boldsymbol{\theta_{im}} \\ \Delta^2\boldsymbol{\theta_{im}} \end{bmatrix} = \begin{bmatrix} \vdots \\ \theta_{im}(2) \\ \theta_{im}(3) \\ \vdots \\ \theta_{im}(n-1) \\ \theta_{im}(n) \\ \theta_{im}(n+1) \\ \vdots \\ \theta_{im}(N-2) \\ \theta_{im}(N-1) \\ \vdots \\ \Delta\theta_{im}(2) \\ \Delta\theta_{im}(3) \\ \vdots \\ \Delta\theta_{im}(n-1) \\ \Delta\theta_{im}(n) \\ \Delta\theta_{im}(n+1) \\ \vdots \\ \Delta\theta_{im}(N-2) \\ \Delta\theta_{im}(N-1) \\ \vdots \\ \Delta^2\theta_{im}(2) \\ \Delta^2\theta_{im}(3) \\ \vdots \\ \Delta^2\theta_{im}(n-1) \\ \Delta^2\theta_{im}(n) \\ \Delta^2\theta_{im}(n+1) \\ \vdots \\ \Delta^2\theta_{im}(N-2) \\ \Delta^2\theta_{im}(N-1) \\ \vdots \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{static mean of the lst aligned model state} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \vdots \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{static mean of the last aligned model state} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{mean delta of the lst aligned model state} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \vdots \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{mean delta of the last aligned model state} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{mean acceleration of the lst aligned model state} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \vdots \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{mean acceleration of the last aligned model state} \\ \\ \\ \end{matrix} \tag{5}$$

respectively, building from the N-by-N linear regression coefficient matrix $\mathbf{W}$.
Consequently, the regenerated LSP trajectories become continuous and smoothly varying, showing similar dynamic change as $\Delta\boldsymbol{\theta}_{im}$ and $\Delta^2\boldsymbol{\theta}_{im}$ described. Before passing to Wiener filtering, the LSPs are converted back to their corresponding LPC coefficients. At this moment, the spectral shape of a target source is ready. The sequence of LPC coefficients $\{a_m\}$ resulted from the regenerated LSP coefficients $\boldsymbol{\theta}'_{im}$ are output to Wiener filtering.

3.2 Wiener filtering and associated LPC formulation

Based on the spectral information of a target source obtained from the preceding trajectory regeneration, Wiener filtering is then used to suppress or to attenuate the interference source. With the LPC coefficients output after trajectory regeneration $\{a_m\}$, the corresponding Wiener filter is derived accordingly. For the present two-source separation problem, the frequency response is

$$H(\omega) = \frac{P_{x,x_i}(\omega)}{P_x(\omega)} \tag{8}$$

where

$$P_{x,x_i}(\omega) = \text{cross power spectral density between } x_1(n) + x_2(n) \text{ and } x_i(n), \quad i \in [1,2] \tag{9}$$

$$P_x(\omega) = \text{auto power spectral density of mixture input } x(n) \tag{10}$$

Assuming that the two source signals have zero crosscorrelation,

$$P_{x,x_i}(\omega) = P_{x_i}(\omega) \tag{11}$$

$$P_x(\omega) = P_{x_1}(\omega) + P_{x_2}(\omega) \tag{12}$$

In terms of LPC coefficients, $H(\omega)$ corresponding to source i becomes

$$H(\omega) = \frac{G_{x_i}^2 B_{x_i}(\omega)}{G_x^2 B_x(\omega)} \tag{13}$$

where G_{x_i} and G_x are the gains of excitation for $x_i(n)$ and $x(n)$, respectively. Furthermore,

$$B_{x_i}(\omega) = \left| \frac{1}{1 + \sum_{m=1}^{M} a_{mx_i} (e^{j\omega})^{-m}} \right|^2 \tag{14}$$

$$B_x(\omega) = \left| \frac{1}{1 + \sum_{m=1}^{M} a_{mx} (e^{j\omega})^{-m}} \right|^2 . \tag{15}$$

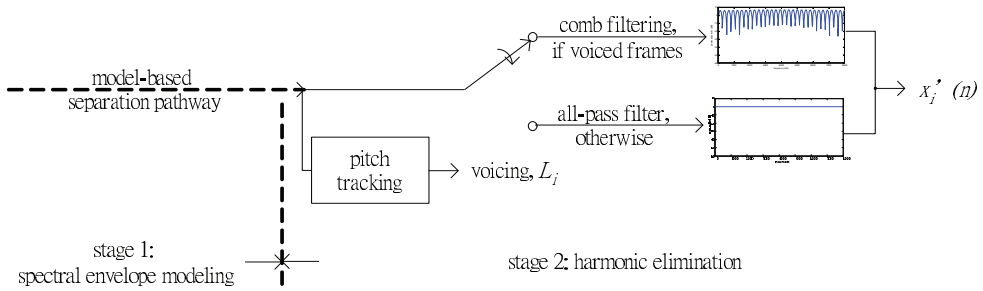


Fig. 10. Block-diagram of harmonic removal of interfering source. For voiced frames, a pitch prediction error filter is used to suppress the harmonics of the interfering source; for unvoiced frames or pauses, an all-pass filter is used instead.

3.3 Harmonic removal of interfering source

A comb filtering in the form of pitch prediction error filter (Ramachandran & Kabal, 1989) is incorporated in the second stage. This stage is used to attenuate harmonic structure associated with the interfering source. For frames where there is no harmonic structure from interfering source, an all-pass filter is applied instead. Fig. 10 depicts the block-diagram. Voicing and pitch information are extracted at the beginning of the second stage.
The impulse response of the pitch prediction error filter $h_p(n)$ is

$$h_p(n) = \delta(n) - \beta_1\delta(n-L) - \beta_2\delta(n-(L+1)) - \beta_3\delta(n-(L+2)) \tag{16}$$

where β_j are the filter coefficients for $j \in [1,2,3]$. L is the lag representing the period of the interfering source.

4. Performance evaluation

Experiments on continuous, real speech utterances are presented below to demonstrate the efficacy of the proposed separation algorithm. It is found that the spectral shape of target speech is retrieved by using the separation algorithm with speech models. Resultant source estimates are close to target speech.
The evaluation set contains 200 mixture signals. These 200 mixture signals are generated by 100 Mandarin source utterances recorded by a female speaker. The signal durations are roughly 3.5 s for each utterance. They are mixed together with equal power, i.e. the signal-to-interference ratio is 0 dB. The evaluation metric used is Itakiura-Saito distortion (d_{IS}).
Itakura-Saito distortion (d_{IS}) concerns about the dissimilarity between a reference speech signal and a test speech signal in an all-pole modeling manner. It represents the degree of spectral matching in terms of general spectral shape and overall gain offset, focusing on the spectral envelope rather than every detail of a speech signal. Fig. 11 depicts the d_{IS} measurements before and after separation.
By comparing the Itakura-Saito distortion values before and after the separation algorithm, the system performance is evaluated. As shown in Fig. 11, most of the IS values of estimated outputs are distributed around IS value = 1 (log value = 0), whereas the IS values of mixture signals are distributed from small to extremely large values. Apart from this, after separation, the mean IS value has been reduced. These confirms that the separation algorithm is effective

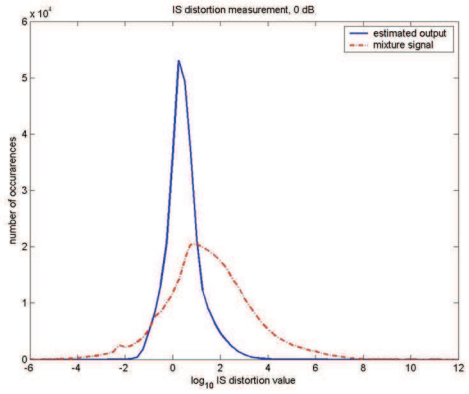


Fig. 11. Measured d_{IS} results before and after separation.

to reduce the Itakura-Saito distortion exhibited in mixture observation and generate source estimate close to target speech source.

Note that there are some extremely small amount of IS values in mixture signals, which probably locate in strong local signal-to-interference regions. After separation, these nearly interference-free regions are filtered and small distortion (as shown by the d_{IS} values after separation) is resulted. In other words, accurate source estimates are achieved.

5. Conclusion

As technologies advance, speech processing applications are no longer limited to controlled usage. More and more applications are designed for daily use, where multiple sound sources with rapidly-changing properties are present. Separation of speech sources become necessary. Over the time, primitive separation cues are often adopted in single-microphone separation algorithms. Nevertheless, schema-based cues are indispensable and critical to separation performance. This chapter first reviews various perceptual cues with the perspective of speech separation. Individual cues are inspected and ranked accordingly to its merits on speech separation. A speech separation algorithm is then introduced to illustrate how primitive and schema-based cues, in particular, in the form of speech models, are incorporated to generate smooth and accurate speech estimates. Experimental results have shown that the effectiveness of this speech model-based separation.

6. References

Assmann, P. F. (1996). Tracking and glimpsing speech in noise: Role of fundamental frequency, *Acoustical Society of America and Acoustical Society of Japan Third Joint Meeting* .

Assmann, P. F. & Paschall, D. D. (1998). Pitches of concurrent vowels, *Journal of the Acoustical Society of America* 103: 1150 – 1160.

Assmann, P. F. & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies, *Journal of the Acoustical Society of America* 88: 680 – 697.

Barker, J. (1998). The relationship between speech perception and auditory organization: Studies with spectrally reduced speech, Ph.D. dissertation, University of Sheffield, August 1998.

Barker, J., Cooke, M. P. & Ellis, D. P. W. (2005). Decoding speech in the presence of other sources, *Speech Communication* 45(1): 5–25.

Barker, J., Coy, A., Ma, N. & Cooke, M. (2006). Recent advances in speech fragment decoding techniques, *Proc. ICSLP*, pp. 85 – 88.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, Cambridge, Massachusetts.

Bregman, A. S., Abramson, J., Doehring, P. & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation, *Perception & Psychophysics* 37: 483 – 493.

Bregman, A. S. & Ahad, P. A. (1996). *Demonstrations of Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, Cambridge, Massachusetts.

Bregman, A. S. & Dannenbring, G. L. (1973). The effect of continuity on auditory stream segregation, *Perception & Psychophysics* 13(2): 308 – 312.

Bregman, A. S., Levitan, R. & Liao, C. (1990). Fusion of auditory components: Effects of the frequency of amplitude modulation, *Perception & Psychophysics* 47(1): 68 – 73.

Bregman, A. S. & Pinker, S. (1978). Auditory streaming and the building of timbre, *Canadian Journal of Psychology* 32: 19 – 31.

Brown, G. J. (1992). *Computational Auditory Scene Analysis: A Representational Approach*, PhD thesis, University of Sheffield.

Brown, G. J. & Cooke, M. (1994). Computational auditory scene analysis, *Computer Speech and Language* 8(4): 297–336.

Brown, G. J. & Cooke, M. P. (1992). A computational model of auditory scene analysis, *Proc. ICSLP*, pp. 523 – 526.

Carlyon, R. P. (1991). Discriminating between coherent and incoherent frequency modulation of complex tones, *Journal of the Acoustical Society of America* 89: 329 – 340.

Carlyon, R. P. (1992). The psychophysics of concurrent sound segregation, *Philosophical Transcactions: Biological Sciences* 336: 347 – 355.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears, *Journal of the Acoustical Society of America* 25: 975 – 979.

Cichocki, A. & Amari, S.-i. (2002). *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, Ltd., Chichester.

Ciocca, V. & Bregman, A. S. (1987). Perceived continuity of gliding and steady-state tones through interrupting noise, *Perception & Psychophysics* 42(5): 476 – 484.

Cooke, M. (1993). *Modelling Auditory Processing and Organisation*, Cambridge University Press.

Cooke, M. P. & Brown, G. J. (1993). Computational auditory scene analysis: Exploiting principles of perceived continuity, *Speech Communication* 13: 391 – 399.

Cooper, L. & Cooper, M. W. (1981). *Introduction to Dynamic Programming*, Pergamon Press Ltd., Oxford.

Dannenbring, G. L. & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex tones, *Perception & Psychophysics* 24(4): 369 – 376.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception, *Journal of the Acoustical Society of America* 76: 1636 – 1647.

Darwin, C. J. (2001). Auditory grouping and attention to speech, *Proc. of the Institute of Acoustics*, Vol. 23, pp. 165 – 172.

Darwin, C. J. & Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution, *Journal of Experimental Psychology: Human Perception and Performance* 3(4): 665 – 672.

Darwin, C. J. & Carlyon, R. P. (1995). Auditory grouping, *in* B. C. J. Moore (ed.), *Hearing*, Academic Press, San Diego, California, chapter 11, pp. 387 – 424.

Darwin, C. J. & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component, *Journal of the Acoustical Society of America* 91: 3381 – 3390.

Darwin, C. J. & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic?, *The Quarterly Journal of Experimental Psychology* 36A(2): 193 – 208.

Deutsch, D. (1999). Grouping mechanisms in music, *in* D. Deutsch (ed.), *The Psychology of Music*, second edn, Academic Press, chapter 9, pp. 299 – 348.

Dowling, W. J. (1973). The perception of interleaved melodies, *Cognitive Psychology* 5: 322 – 337.

Ellis, D. P. W. (1994). A computer implementation of psychoacoustic grouping rules, *Proc. 12th International Conference on Pattern Recognition*, Vol. 3, pp. 108 – 112.

Ellis, D. P. W. (1996). *Prediction-driven Computational Auditory Scene Analysis*, PhD thesis, Massachusetts Institute of Technology.

Flanagan, J. (1972). *Speech, Analysis, Synthesis and Perception*, 2nd ed. Springer-Verlag, Berlin.

Gardner, R. B., Gaskill, S. A. & Darwin, C. J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency, *Journal of the Acoustical Society of America* 85: 1329 – 1337.

Hall, J. W., Haggard, M. P. & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis, *Journal of the Acoustical Society of America* 76: 50 – 58.

Helmholtz, H. (1954). *On the sensations of tone as a physiological basis for the theory of music*, 2nd ed. Dover Publications, New York.

Hu, G. & Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Trans. Neural Networks* 15(5): 1135–1150.

Hu, G. & Wang, D. (2006). An auditory scene analysis approach to monaural speech segregation, *in* E. Hänsler & G. Schmidt (eds), *Topics in Acoustic Echo and Noise Control*, Springer, chapter 12, pp. 485 – 515.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals, *Journal of the Acoustical Society of America* 57: S35.

Jansson E. V. (1978). Tone characteristics of the violin, *Svensk Tidskrift för Musikforskning (Swedish Journal of Musicology)* STM 1978:1: 83–105.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts.

Klapuri, A. P. (2001). Multipitch estimation and sound separation by the spectral smoothness principle, *Proc. ICASSP*, pp. 3381 – 3384.

Koffka, K. (1963). *Principles of Gestalt Psychology*, Harcourt, Brace & World, Inc., New York.

Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes, *in* M. Kubovy & J. R. Pomerantz (eds), *Perceptual Organization*, Lawrence Erlbaum Associates, Publishers., chapter 3, pp. 55 – 98.

Lee, S. W., Soong, F. K. & Ching, P. C. (2006). An iterative trajectory regeneration algorithm for separating mixed speech sources, *Proc. ICASSP*, pp. I–157 – I–160.

Lee, S. W., Soong, F. K. & Ching, P. C. (2007). Model-based speech separation with single-microphone input, *Proc. Interspeech*, pp. 850 – 853.

Lieberman P. (1961). Perturbations in vocal pitch, *Journal of the Acoustical Society of America* 33: 597 – 603.

Lim, J. S. (1983). *Speech Enhancement*, Prentice Hall, Englewood Cliffs, New Jersey.

Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton.

Marin, C. M. H. & McAdams, S. (1991). Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width, *Journal of the Acoustical Society of America* 89: 341 – 351.

McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence, *Journal of the Acoustical Society of America* 86: 2148 – 2159.

Meddis, R. & Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies, *Journal of the Acoustical Society of America* 91: 233 – 245.

Mellinger, D. K. (1991). *Event Formation and Separation in Musical Sound*, PhD thesis, Stanford University.

Paliwal, K. K. (1989). A study of line spectrum pair frequencies for vowel recognition, *Speech Communication* 8: 27 – 33.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection, *Journal of the Acoustical Society of America* 60: 911 – 918.

Purwins, H., Blankertz, B. & Obermayer, K. (2000). Computing auditory perception, *Organised Sound* 5: 159 – 171.

Rabiner, L. & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey.

Ramachandran, R. P. & Kabal, P. (1989). Pitch prediction filters in speech coding, *IEEE Trans. Acoust., Speech, Signal Processing* 37: 467 – 478.

Scheffers, M. T. M. (1983). *Sifting Vowels: Auditory Pitch Analysis and Sound Segregation*, PhD thesis, University of Gröningen, The Netherlands.

Soong, F. K. & Juang, B.-H. (1984). Line spectrum pair (LSP) and speech data compression, *Proc. ICASSP*, pp. 37 – 40.

Summerfield, Q. & Culling, J. F. (1992). Auditory segregation of competing voices: Absence of effects of FM or AM coherence, *Philosophical Transcactions: Biological Sciences* 336: 357 – 366.

van der Kouwe, A. J. W., Wang, D. & Brown, G. J. (2001). A comparison of auditory and blind separation techniques for speech segregation, *IEEE Trans. on Speech and Audio Processing* 9: 189–195.

Viterbi, A. J. (2006). A personal history of the viterbi algorithm, *IEEE Signal Processing Magazine* 23: 120 – 142.

Wakita, H. (1981). Linear prediction voice synthesizers: Line-spectrum pair (LSP) is the newest of several techniques, *Speech Technology* pp. 17 – 22.

Wang, D. & Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, John Wiley & Sons, Ltd., Hoboken, New Jersey.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds, *Science* 167: 392 – 393.

Warren, R. M. (1982). *Auditory Perception: A New Synthesis*, Pergamon Press, Elmsford, New York.

Warren, R. M. & Obusek, C. J. (1971). Speech perception and phonemic restorations, *Perception & Psychophysics* 9: 358 – 362.

Warren, R. M., Obusek, C. J. & Ackroff, J. M. (1972). Auditory induction: Perceptal synthesis of absent sounds, *Science* 176: 1149 – 1151.

Weintraub, M. (1985). *A Theory and Computational Tool of Auditory Monaural Sound Separation*, PhD thesis, Stanford University.

Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

S. W. Lee (2011). Single-Microphone Speech Separation: The use of Speech Models, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, Available from: http://www.intechopen.com/books/speech-technologies/single-microphone-speech-separation-the-use-of-speech-models

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.