On noise masking for automatic missing data speech recognition: A survey and discussion

Christophe Cerisara *, Sébastien Demange, Jean-Paul Haton

LORIA, UMR 7503, Nancy, France

Received 29 September 2005; received in revised form 21 July 2006; accepted 28 August 2006
Available online 28 September 2006

Abstract

Automatic speech recognition (ASR) has reached very high levels of performance in controlled situations. However, the performance degrades significantly when environmental noise occurs during the recognition process. Nowadays, the major challenge is to reach a good robustness to adverse conditions, so that automatic speech recognizers can be used in real situations. Missing data theory is a very attractive and promising approach. Unlike other denoising methods, missing data recognition does not match the whole data with the acoustic models, but instead considers part of the signal as missing, i.e. corrupted by noise. While speech recognition with missing data can be handled efficiently by methods such as data imputation or marginalization, accurately identifying missing parts (also called masks) remains a very challenging task. This paper reviews the main approaches that have been proposed to address this problem. The objective of this study is to identify the mask estimation methods that have been proposed so far, and to open this domain up to other related research, which could be adapted to overcome this difficult challenge. In order to restrict the range of methods, only the techniques using a single microphone are considered.
© 2006 Elsevier Ltd. All rights reserved.

1. Introduction

Automatic speech recognition has reached very high levels of performance in controlled situations, when the speaker is known a priori and when background noise is limited. However, nowadays the accuracy of speech recognition systems dramatically degrades when the background noise increases. Many approaches have been proposed to improve the robustness to noise, but very few of them can handle highly non-stationary noise, such as background speech or music. Among this limited class of algorithms, missing data recognition has shown very promising results. The main idea of missing data recognition is to explain only a sub-part of the observed signal with stochastic models, and to “ignore” the noisiest (or less important) observations. A similar principle is used in a few other robust methods, such as uncertainty decoding or detection-based recognition, but the main characteristic that sets missing data recognition apart from these methods is the use of masks. These so-called missing data masks play a fundamental role in missing data recognition, and their quality has a strong

* Corresponding author.
E-mail address: cerisara@loria.fr (C. Cerisara).

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.08.001
impact on the final recognition accuracy. This paper is thus dedicated to mask estimation for missing data recognition. Its objective is to review all the possible approaches that can be used to identify the masks. These approaches are classified into two main classes: first, methods that mimic, or are inspired by the human ear, and second, techniques that are based on signal processing and stochastic modelling. Algorithms that exploit two or more microphones are not considered in this review. The localization cue that can be exploited in such cases is very important, and has also proven many times to be very effective. However, this paper focuses exclusively on single-channel approaches, since this situation is the most commonly encountered in practical ASR applications.

This paper is organized as follows: Section 2 defines the missing data masks, and discusses their importance and role in missing data recognition. Section 3 reviews the methods based on Computational Auditory Scene Analysis (CASA). The main architectural solutions that have been proposed for CASA systems are described. Section 4 focuses on various approaches that are not closely related to CASA. It begins with approaches based on noise estimates, and shows how the mask models can be trained and possibly combined with other (speech) models. Alternative techniques based on Independent Component Analysis (ICA) or neural networks are also presented. Section 5 discusses the possible evaluation criteria and future works, and Section 6 concludes the paper.

2. Masking speech signal

In this section, missing data masks are defined, and their importance and role in robust automatic speech recognition analyzed.

2.1. Definition of masks

In the area of speech recognition, the term “mask” can have different meanings depending on the task and objective. Before detailing these possible definitions, it should be noted that masking is very common in the human auditory process. For example, the following masking schemes have been identified (Moore, 1982):

- The “capture” effect, when locally more intense sound components dominate the neural response, or when the mask frequency is close to the frequency of the signal.
- Central masking when the mask and the signal are presented to different ears.
- Forward and backward masking when the mask is presented before or after the signal.

In the context of automatic missing data recognition, masks can be defined as a function $M$ that associates a value in a domain $D$ to any speech coefficient. Although in theory a mask can be defined for any parameterization domain, in practice, such a domain should map distinct frequency bands onto different feature space dimensions, so that a frequency-limited noise only affects a few dimensions of the feature space. This is typically true for frequency-like and wavelet-based parameterizations and for most auditory-inspired analysis front-ends. This is not the case in the cepstrum. In the following, we will consider without loss of generality the frequency domain.

We assume next that the time and frequency dimensions are both discrete, which means that any speech coefficient is represented by a couple of positive integers. $M$ is thus defined as

$$M : (\mathbb{N}, \mathbb{N}, \Omega) \rightarrow D$$

$M$ further depends on a domain $\Omega$ that contains all the other information required to compute the mask.

Let us first consider the mask domain $D$. Two solutions are commonly employed in the missing data community:

- $D = \{0, 1\}$: Such masks are usually called hard masks: any coefficient is either masked or not.
- $D = [0, 1]$: Such masks are usually called soft masks. Their value is interpreted as the probability that the coefficient is masked. Until now, its only interpretation is probabilistic, but it would also be interesting to investigate alternative interpretations, such as a possibility of fuzzy logic.
Several missing data studies have shown that soft masks give better results than hard masks, at least for the most common definitions and uses of masking in speech recognition.

2.2. Mask decision domain

Let us now consider the “decision” domain $\Omega$. A coefficient is usually masked when, for this coefficient, the contribution of the useful signal is dominated by the contribution of the noise. Such a criterion can be interpreted as applying a threshold to the local signal-to-noise ratio (SNR).

When $\Omega$ contains the exact value of the local SNR, which can be computed given both the clean and noisy speech signals, the resulting mask is called an oracle mask. Oracle masks have often been used in the literature to estimate upper-bound limits on the recognition accuracy that can be obtained using missing data techniques. Obviously, in real situations, the exact value of the local SNR is not known. The crucial point is how to estimate this signal-to-noise ratio, and the purpose of this paper is to review the most important approaches to making that estimation.

Sometimes $\Omega$ is not limited to the local SNR, but has additional constraints, such as minimum duration and bandwidth for masked fragments, or smoothness of the masked values over time and frequency. These constraints may improve the SNR-estimate, produce masks that can be more easily interpreted by humans, and further considerably reduce the search space of plausible masks. A particular case is to constrain all the coefficients of the same frame to share the same masking value. A single mask is then defined for the whole frame, rather than the coefficient. This particular case is related to the usable speech research area, where the frames that do not carry useful information (possibly because they are noisy) are discarded. These global masks are also used in detection-based algorithms, such as in Wang (2004). Although some mask estimation techniques that come from these research areas will be cited later on in this review, the general focus of this paper is on estimating one mask per coefficient.

The SNR-based definition of $\Omega$ is the most intuitive, and the most widely accepted in the speech recognition community, but other definitions can be considered, depending on the speech-processing task. When the final objective is to denoise the signal in order to increase its SNR, then this definition is quite appropriate. But real applications often aim instead at improving the intelligibility of the speech signal, in which case an increase in the SNR does not systematically translate into an increase in speech intelligibility. This is also true when the objective is to automatically recognize the speech signal. In such cases, it might be better to base the decision on word accuracy rather than on local SNR. However, to the best of our knowledge, no algorithms to compute such accuracy-based masks have been published so far. There are two main reasons for this:

- SNR-based masks are much easier to compute than masks that maximise the recognition performance.
- Very good results have been reported with SNR-based masks and missing data recognition, which suggests that such masks are already good approximations of word accuracy-based masks.

Nevertheless, it might be interesting in the near future to investigate new kinds of masks that directly optimize the recognition accuracy, and to train stochastic models to infer such masks in unknown conditions.

2.3. On the use of masks

There are two common techniques for handling missing features during recognition. The first involves replacing all the missing values by their estimated value in order to provide the recognizer engine with a completely observed signal. This approach is called data imputation. Several methods have been proposed to infer the missing values in the framework of data imputation. The most popular ones are presented and compared in Raj (2000). The second technique is called data marginalization: the recognizer engine is modified in order to handle partial feature vectors. The observation probability is computed by integrating the likelihood of the missing features over the whole range of possible values, while reliable features are processed as usual. Several algorithms for data marginalization have also been proposed. These variants depend on the information available about the true value of missing data (Barker et al., 2000; Morris, 2001). Experimental results for both imputation and marginalization are given in Morris et al. (2001).
2.4. Mask estimation

Computing missing data masks can be realized either independently from, or jointly with the speech decoding process. The first missing data recognition methods were based on masks that were pre-computed independently from the speech decoding process using signal-processing techniques. This is the case for example of the CASA-based approaches that are presented in Section 3. More recently, authors have shown that more efficient masks can be derived when these masks are estimated jointly with the speech decoding hypothesis, as in Barker et al. (2001a) or Roweis (2003). In both cases, masks still have a fundamental role to play regarding the robustness of speech recognition. In fact, the main difference between missing data approaches and other robust methods, such as uncertainty decoding, is that missing data recognition makes use of masks. Basically, uncertainty decoding assumes that the noisy speech \( y \) is observed, while the clean speech \( x \) is hidden. Given the clean speech models \( \lambda \), the observation likelihood is then:

\[
p(y|\lambda) = \int p(y|x, \lambda)p(x|\lambda)dx
\]

\( p(y|x, \lambda) \) is modelled by a probability density function in uncertainty decoding. Missing data marginalization follows exactly the same principle, except that it assumes a different probability density function, depending on whether a given coefficient is masked or not: \( p(y = x|\lambda) = 1 \) for unmasked coefficients, and \( p(y|x, \lambda) \) is uniform for masked coefficients. Masks thus play a central role in missing data recognition.

Furthermore, even though masks can be entirely derived from speech models in specific situations, as in Roweis (2003), we are convinced that the general problem of noise robust automatic speech recognition can only be solved by exploiting and merging multiple sources of information other than speech and noise models, such as cues derived from a direct analysis of the signal, or from the knowledge that experts may have on speech production, phonetics, psycho-acoustic, etc. All these types of information that cannot be modelled stochastically should also be considered when computing missing data masks.

This analysis leads to another (classical) clustering of mask estimation methods into bottom-up vs. top-down techniques. Bottom-up approaches compute deterministic properties of the speech signal to analyze it and derive the masks, while top-down methods exploit mask models. There is now a consensus in the missing data recognition community to combine both techniques.

3. CASA-based approaches

Auditory Scene Analysis (ASA) refers to the ability of humans to analyse their acoustic environment in order to identify every single object and entity that contributes to the perceived signal and to focus his attention on one or several of these sound sources of interest. Computational Auditory Scene Analysis (CASA) aims at developing automatic models and algorithms that can achieve the same objective, i.e. identify every sound source, extract the audio stream(s) of one or several of these sound sources, or focus on one of these streams in order to process it (event detection, speaker recognition, ...). In the context of missing data recognition, the role of CASA approaches is limited to estimating the audio masks, i.e. segregating the regions of the spectro-temporal domain that are dominated by the speech of the target speaker from the ones that belong to any of the other (noise) sound sources.

3.1. CASA basic principles

A first approach to estimating acoustic masks is to mimic the human ear in its ability to focus on a single target speaker, even in severe noisy conditions like at a cocktail party. Many studies have tried to understand how the human ear achieves this robustness: see (Bronkhorst, 2000) or (Cooke and Ellis, 2001) for a review. Bregman’s book (Bregman, 1990) defines the fundamental CASA principles on which most CASA systems are based.

Bregman noticed that by editing some features of speech sounds, these sounds could be perceived as either one sound or as several sounds. This shows that a process exists in the human brain of collecting the sounds and perhaps merging them. His theory proposes two classes of organization. The first one is simultaneous...
fusion: for example, sinusoids that are harmonically related are fused into a single tone. Detuning one of the harmonics, or delaying its onset or offset, can cause it to be perceived as a separate sound. Simultaneous onset is particularly important. Two other cues for fusion are common spatial location and modulation. The second class is sequential organization: successive events are assigned to separate streams according to principles derived from Gestalt psychology (Handel, 1986; Miller and Heise, 1950). Bregman suggests that such a primitive organization is pre-wired in the auditory apparatus. He further proposes another grouping/segregation process that is driven by learned schemas. While primitive grouping does not depend on the sound source, schemas are specific to a type of sound source, such as speech for example.

3.2. A brief overview of CASA systems

The most important CASA systems are reviewed in Brown, and Wang (2005). The first CASA systems (Weintraub, 1985; Brown and Cooke, 1994) were merely based on a bottom-up approach, in which features are computed to group or segregate the contributions of different sound sources. For example, Brown’s system builds onset and offset maps, which are based on the leaky sum of two excitatory and inhibitory inputs, as modelled in Shofner and Young (1984). It also exploits autocorrelation and frequency transition maps to group channels that are responding to the same acoustic component. Another bottom-up CASA system that exploits continuity of timbre has been proposed in Masuda-Katsuse and Kawahara (1999). Masuda-Katsuse et al. show that the continuity of spectral change is an important criterion that is used in the sequential integration process depicted by Bregman. Although the existence of such a criterion was denied by Bregman, the authors of (Masuda-Katsuse and Kawahara, 1999) report psychophysical experiments that confirm its existence. Their system further exploits harmonicity to group the frequencies of concurrent sound streams. Then, sequential grouping of the streams is based on spectral tracking, which is implemented as a second-order auto-regressive state-space model. Amplitude modulation can also be used to detect missing data masks in the high-frequency parts of the spectrum (Hu and Wang, 2004): put simply, high-frequency filters have a wide bandwidth and respond to multiple harmonics, which create beats and combinational tones that modulate their amplitude.

The design of a more generic computational approach to handling these fusion and segregation procedures was not addressed in these systems. Three main computational models have been proposed, respectively based on the blackboard, multi-agent and neural-network architectures. These three approaches are briefly reviewed below.

3.2.1. Blackboard CASA systems

Godsmark and Brown (1999) base their research on the observation that the organisation in the human ear is context-sensitive and retroactive, which suggests that the decision of contradictory interpretations of the signal may be postponed with the expectation that disambiguating evidence will emerge later. They thus propose letting the grouping principles interact within a sliding window of 300 ms, and then imposing a fixed organisation at the end of this window.

The system has been validated on musical signals. It is based on a blackboard architecture, decomposed into eight stages, in bottom-up order:

(1) Parameterization of the signal into synchrony strands (Cooke, 1993), each of which represents a dominant spectral component that extends through time and frequency. From one analysis window to the next, synchrony strands are extended based on temporal continuity, frequency proximity and amplitude coherence: the latter principle is used to prevent different musical instruments that would play harmonic sounds from being grouped together.

(2) Feature strands are built from the set of synchrony strands. They include onset and offset time, initial and final frequency, and a frequency transition history.

(3) Synchrony strands that are likely to come from the same sound source are grouped together. Several experts fire grouping hypotheses on the blackboard, based on onset and offset synchrony, temporal and frequency proximity, harmonicity and common frequency movement. Hypotheses at this stage correspond to notes.
Hypotheses are evaluated by experts and sorted in order of decreasing likelihood. The evaluation measures how well a hypothesis fits certain principles, such as time–frequency proximity or onset synchrony.

Emergent properties are derived from grouped strands, i.e., fundamental frequency and timbre.

A new grouping of strands is performed based on these emergent properties. These new groups correspond to melodic lines. The principles used are pitch proximity and timbral similarity. Timbre is modelled as changes in the spectral centroid, which can be interpreted as a “brightness envelope”, and the amplitude envelope.

These new hypotheses are also evaluated.

At the highest level, source-specific knowledge is used to identify meter and repeated melodic phrases.

Ellis et al. (1996) proposes a prediction-driven blackboard architecture to implement a CASA system, which is based on internal world models that are composed of objects whose aggregate explains and predicts the observed signal. These models are used to predict the next occurring sound, which gives an initial hypothesis that can serve as a starting point for searching for a better hypothesis. The world models contain three classes of sound: noise clouds, tonal elements, and transients. An abstraction hierarchy is based on these sound elements to build up structured patterns of sounds. From the current features of the signal (such as energy envelope), the next set of values is predicted based on the models that have been selected at the previous steps. These predictions are probabilistic. When the next observed set of parameters is available, it is compared with the prediction: if it fits within the margin of error of the prediction, the activated models are simply updated with these new values, otherwise, new models will be activated, or old models deactivated, to account for the under- or over-estimated energy.

3.2.2. Agent-based CASA

Nakatani (2002) proposes another interesting architecture for CASA. This agent-based architecture is very flexible. The front-end is also based on harmonic features. Generator agents detect the beginning of harmonic structures and generate a new tracer agent. This is realized via pitch watcher agents, each of which evaluates the harmonic intensity at one possible fundamental frequency. The pitch watcher with the highest harmonic energy generates a tracer agent. The tracer agent searches for the value of the fundamental frequency closest to the previous detected value. Each time a tracer agent is activated, it subtracts its own harmonic structure from the signal: the residual signal can then be processed by other tracers and by the generator. Monitor agents are further introduced to detect and eliminate redundant and ghost tracers.

Such a sub-system that extracts sound streams using certain sound attributes is called an agency. Other agencies are proposed, which respectively exploit visual cues, spatial processing, and binaural or monaural processing. Each agency produces its own set of streams, and different kinds of interaction between agencies may occur: for example, the outputs of two monaural agencies may be the input of a binaural agency, while an agency that detects voiced speech fragments may replace or suppress the output of the binaural agency. To illustrate these interactions between different agencies, the authors extend their system to take into account both harmonicity and spatial localisation sound segregations.

3.2.3. Neural-based CASA systems

Wang and Brown (1999) developed a two-layer oscillator network for sound stream segregation. Each auditory feature is encoded into relaxation oscillators. Synchronized oscillators are merged into a single stream, while desynchronized ones are assumed to represent different streams. This model improves the signal-to-noise ratio of a speech signal mixed with different background noise types, but does not outperform blind source separation when multiple interfering sources are added (van der Kouwe et al., 2001) see Fig. 1.

Another neural architecture, the cortronic network (Sagi et al., 2001), was proposed for word recognition in background noise or with multiple speech interferences. It is based on an associative memory neural network made up of three layers: one for sound encoding, one for sound processing and the third one for word processing.

Several important concepts about the implementation of CASA with neural networks are discussed in Haykin and Chen (2005). They propose an architecture that exploits a novel idea, i.e. active audition, in order to address the cocktail-party problem. This architecture is based on four layers:
• Localization, which infers the directions of incoming signals. This can be realized with an array of microphones.
• Segregation and focal attention, which segregates the different sound sources and focuses on one of them.
• Tracking, which tracks the evolution of the signal features and predicts their future value.
• Learning, which continuously updates the internal model to adapt it to the evolution of the perceived world.

4. Solutions based on signal processing and statistical models

The mask estimation methods that are reviewed in this section do not claim to copy Bregman’s principles, nor to be strongly inspired by the human auditory system. Instead, they use signal-processing concepts and statistical approaches to identify the masks.

4.1. SNR-based segregation

As discussed in Section 2, common missing data are directly related to the local signal-to-noise ratio (SNR). Hence, most missing data recognition studies somehow estimate the local SNR to build these masks. Virtually any noise estimation algorithm, amongst the numerous ones that have been published, can be used for that purpose. The simplest estimates the local SNR from a noise model, usually computed on segments of the signal without speech energy, in a similar way to spectral subtraction (Cooke et al., 1997; Renevey, 2001; Renevey and Drygajlo, 2001; Morris et al., 2001). However, this simple solution suffers from its lack of robustness against non-stationary noise, which originally motivated the missing data recognition framework. Therefore, more sophisticated algorithms are used to improve the local SNR estimation, for example the Vector Taylor Series method in Raj (2000). Dupont and Ris (2001) assess and compare four methods for the local estimation of noise spectra, namely energy clustering, the Hirsh histograms, the weighted average and low-energy envelope tracking. In addition, pseudo-code for these four algorithms is given. Although these techniques improve the noise level estimation accuracy, gain is not significant enough when the corrupting noise is non-stationary. Additional signal-processing cues, such as harmonicity and amplitude modulation, are then often computed and combined with noise models to improve local SNR estimation. For the particular case of reverberant speech, (Palomaki et al., 2004) proposes to exploit modulation filtering to detect strong speech onsets that have not been contaminated by distortion and to generate a missing data mask from these cues.

Harmonicity measures have been investigated to compute missing data masks for the frame or the coefficient. For the frame, a related research area, called usable speech, aims at identifying and extracting those portions of co-channel speech that are still useful for speech-processing applications such as speaker identification or speech recognition (Yantorno et al., 2003). In this context, the two following measures have been proposed:
• Adjacent Pitch Period Comparison, which compares the Euclidean distance between adjacent pitch periods of voiced co-channel speech. When there is little interference, this distance is usually short, and conversely it is long when there is substantial interference.

• Spectral Autocorrelation Peak Valley Ratio-Residual (SAPVR), which detects usable speech by examining the structure in the autocorrelation of the FFT of the Linear Predictive Coding (LPC) residual (Chandra and Yantorno, 2002). When there is little interference, a pattern of peaks and troughs occurs, and conversely, when the interference is substantial such patterns fail to appear. The success of SAPVR-Residual measure depends highly on the pattern of peaks and troughs, so it is essential to have a robust peak-picking algorithm.

In Chandra and Yantorno (2002), both measures are merged. ICA is first applied to both measures to eliminate redundant information, which improves the quality of the fusion Hall (1992). Fusion is realized by a third-order least square error (LSE) system, and also by a Bayesian classifier.

When harmonicity measures are applied to each coefficient, the basic hypothesis is that the harmonic overtones of each speech source can be grouped together into the same acoustic stream (de Cheveigne et al., 1995). The frequency bands that are not part of this stream have a higher probability of being masked. Hence, (Barber et al., 2001b) combined a harmonicity mask with a traditional SNR-based mask. Their harmonicity mask is obtained by taking a slice through the autocorrelogram at the lag corresponding to the pitch period. This slice is then rescaled using a sigmoid function to obtain a frame of the soft harmonicity mask. The combined mask is then a weighted sum of the harmonicity and SNR masks. Weights correspond to degrees of voicing of each frame rescaled using a sigmoid function. A number of other missing data works, such as (van Hamme, 2004), also combine harmonicity and SNR for mask estimation.

A more difficult case occurs when the noise is also harmonic: multi-pitch tracking algorithms have to be used to distinguish the concurrent harmonic scales (de Cheveigne, 1993; Parsons, 1976). Sinusoidal modelling can be useful for this purpose (Tolonen and Karjalainen, 2000; Karjalainen and Tolonen, 1999): once the harmonic signal is decomposed into its sinusoidal components, these components can be grouped together into one or several sound sources based on their frequency ratio. (Virtanen and Klapuri, 2002) extend this principle and integrate it into a complete sound separation system. Alternatively, the main pitch period can be detected, and the corresponding harmonics removed from the signal. A new pitch estimate can then be iteratively computed on the residue of the signal (Nakatani, 2002).

Tchorz and Kollmeier (2002) propose an original method for estimating the local SNR. Motivated by neurophysiological findings on amplitude modulation processing in higher stages of the auditory system in mammals, their method analyses information on both centre frequencies and amplitude modulations of the input signal. This information is represented in two-dimensional, so-called amplitude modulation spectrograms (AMS). Then, a neural network is trained on a large number of AMS generated from mixtures of speech and noise. An SNR estimation for the frame is given at the output of the neural network. Experiments show that the proposed approach functions well even if the corrupting noise is non-stationary. The authors further propose extending this algorithm to estimate the local SNR in different frequency channels (Tchorz et al., 2001). Missing data masks could then be computed from the outputs of each channel-specific neural network.

4.2. Statistical models of the masks

In the previous sections, bottom-up mask estimation methods have been presented. In this section, top-down approaches based on mask models are reviewed, while in the next section, we consider the case where mask estimation also depends on the speech models.

The most intuitive Bayesian approach to solving the mask estimation challenge is to model the following probability density function:

\[ p(Y_t|m_{i,j}) \]

where \( m_{i,j} \) indicates whether the corresponding pixel is masked or not, and \( Y_t \) is a set of features computed from the observed signal. Any standard classifier can be used for this purpose: neural networks, Gaussian mixture models, support vector machines, etc.
This is the approach followed by Seltzer (2000) and Raj (2000). Raj builds a Bayesian classifier for each frequency band, and computes a feature vector that is composed of the spectral energy in the band along with its gradient in time and frequency. Seltzer computes features that are designed to represent the speech characteristics. He proposed the following:

- the comb-filter ratio, which represents the energy in the harmonics, an alternative to the harmonicity mask of Barker Barker et al. (2001a);
- the autocorrelation peak ratio, which measures the degree of periodicity of the signal, and is thus related to the SNR, when we assume that the noise is much less harmonic than speech;
- the sub-band energy to full-band energy ratio, which represents the overall shape of the spectrum of speech;
- the Kurtosis, which measures the Gaussianity of the signal (clean speech signals have a higher kurtosis than noisy speech signals);
- the flatness of the spectral valleys, which is related to the SNR;
- the sub-band energy to sub-band noise floor ratio, which is based on an estimate of the noise floor energy;
- a spectral-subtraction-based SNR estimate.

In the work of Seltzer, white noise is used for training the classifier for mask estimation in order to reduce the effects of unknown noise. (Kim et al., 2005) note that training the Bayesian mask classifier on white noise cannot provide a good recognition accuracy over all types of noise. They argue that white noise should be predictive of results for other types of noise if the mask estimates obtained in every sub-band are independent from those in adjacent frames and other sub-bands. Unfortunately, spectral values do not follow these requirements. Therefore, (Kim et al., 2005) propose incorporating the spectral variations across frames and sub-bands into each sub-band model by training them on speech databases that have been corrupted by various random combinations of colored noise with random durations. Their experimental results show that the proposed training method is encouraging. This approach is further improved in Kim and Stern (2006), in particular by training band-independent mask models.

### 4.3. Masks estimated during the recognition process

The multi-source decoder (Barker et al., 2001a) explicitly combines primitive grouping and a model-driven architecture. Basically, primitive grouping principles are used to group spectro-temporal regions likely to belong to the same sound source into fragments. Then, acoustic speech models act as schemas in the second stage of Bregman's model to select and group together the fragments that might come from the same speaker. This system relates to the ASA model called "glimpsing model" (Cooke, 2003a; Cooke, 2003b), in which speech hearing exploits localized regions of the time–frequency plane, or glimpses of speech, when such glimpses dominate the spectrum for a minimum duration and within a continuous frequency range of a minimum bandwidth. Detailed evidence for this model is given in Barker et al. (2005), along with a first implementation in the context of missing data recognition. This implementation is also based on both primitive grouping and top-down models of speech. This combination of bottom-up and top-down approaches is a very important requirement for handling background speech noise, as discussed in Remez et al. (1994), Barker and Cooke (1999) and Cooke (2006), where sine-wave target and masker speech intelligibility experiments are carried out, with the conclusion that primitive grouping alone cannot explain the human performance obtained in such conditions.

When two models for the target speech and the background noise (or speech) are available, it is possible to combine both these models and to search for the best path that maximizes the observation likelihood in the combined state space. From this point of view, (Roweis, 2003) proposed a method based on sub-band processing to separate two speakers talking simultaneously. He assumes that only one speaker dominates each time–frequency pixel. He then uses a combination of the speaker-specific models to find the best possible alignment.

A related work is proposed in Reyes-Gomez et al. (2004). It is inspired from the vision domain, where several layers are used to account for different sources of variability. For speech, two layers are defined: one for the excitation harmonics, and another for the resonances. The authors compute a transformation of the cur-
rent frame to predict the following one. Transformations are encoded in a generative graphical model that relates the excitation harmonics $F$, the filter (formants) $H$, their transformations $T$ and the spectrum $X$. Source separation can then be realized by building one layer per speaker.

Another approach is proposed in Kristjansson et al. (2004), where the simultaneous speech of two speakers are segregated from their estimated posteriors $p(x_1|y)$ and $p(x_2|y)$. The combination of both signals is modelled by a normal density function in the log-spectral domain $p(y|x_1,x_2)$. This combination is modelled by a random process, because of the angle between the complex spectra of both speakers, which gives an additive term that is modelled as a random “error” variable in the log-spectral magnitude domain. The mean of this normal density is then linearized with respect to $x_1$ and $x_2$, in order to approximate the posterior $p(x_1,x_2|y)$ by a Gaussian. The linearization point is iteratively adjusted towards the mode of the true posterior. Two GMMs, one for the male $p(x_1)$ and another for the female speaker $p(x_2)$ are trained, and finally used to segregate both speaker streams.

### 4.4. One-microphone source separation

We have seen in Section 2.2 that oracle missing data masks can be computed from segregated speech and noise streams. An easy method of computing missing data masks is to use exactly the same procedure, but with an estimate of both speech and noise streams. In other words, source-separation algorithms can be used as a pre-processing step before estimating missing data masks, in a similar way to how CASA systems have been combined with missing data recognition techniques. These source-separation algorithms output a speech stream that very often still contains remaining noise or distortion. Hence, directly recognizing such streams often leads to poor recognition rates, and it is usually better to build missing data masks from these streams and subsequently apply missing data recognition techniques.

A number of research areas propose sound source-separation algorithms, including CASA, blind source separation and co-channel speech separation (Quatieri and Danisewicz, 1990; Yen and Zhao, 1999). Many blind source-separation algorithms exploit several microphones. When only one microphone is considered, (Potamitis et al., 2001) propose the application of unifying Bayesian Independent Component Analysis of two different methods to denoise speech signals. The first method is based on Sparse Code Shrinkage, which computes the unmixing matrix $W$ from a large set of clean speech frames. For testing, $W$ is applied to the noisy frames. The hidden clean speech components are independent, and can be factorized in terms of pdf marginals. The posterior pdf of the clean frames given the noisy ones is factorized into its clean and noisy components, and is used to derive the maximum a posteriori (MAP) estimate of the clean frames only. The second method exploits Variational Bayes approximation: the mixing matrix is now different for each frame and is estimated from the noisy signal. A mixture of Gaussian priors is assigned to the sources. The model parameters are inferred by a free-form functional minimization of the Kullback–Leibler divergence between the approximate distribution of the posterior distribution and the true posterior.

Instead of ICA, (Bach and Jordan, 2005) exploit spectral learning to train a segmenter that segregates all the points in the spectrum into two subsets, one per speaker. This segmenter takes the form of parameterized affinity matrices, which encode topological relationships in the spectrum for given features. The features computed are continuity in time and frequency, common time variations, harmonicity and timbre. The parameters...
of the segmenter are trained on the mixed speech from a small set of speakers, and are tested on the mixed speech of different speakers. The main advantage of this approach is that it does not require precise models of the mixed voices.

4.5. Neural networks

Potamitis et al. (2000a) exploit neural networks to detect usable speech frames corrupted by impulsive noise. Then, the corrupted vectors are restored using a second, independent neural network. The authors have extended this technique to detect and handle missing data masks in the context of missing data recognition (Potamitis et al., 2000b). The whole spectrum is divided into frequency bands, and a Time Delay Neural Network (TDNN) is trained on the clean speech derived from each band to predict the value of the next coefficient. The prediction error realized by the TDNN is evaluated on a development corpus. During testing, the observed coefficients that are outside the 99% statistical control limits are masked.

5. Discussion

5.1. Evaluation

The evaluation criterion for systems that compute and exploit missing data masks clearly depends on the target application. In the context of this paper, the target application is speech recognition, and the preferred evaluation criterion is thus word recognition accuracy. All the methods that have been reviewed in this paper can be integrated within a missing data speech recognizer. However, many of these works do not actually achieve this integration or do not compute recognition accuracy.

For this reason, CASA systems are often evaluated on source-separation tasks, by computing SNR improvements (Cooke, 1993). This is the case for example of (Brown and Cooke, 1994). Similarly, (Wang and Brown, 1999 and Hu and Wang, 2004) are both evaluated based on SNR-reduction, and compared with each other. Some CASA systems are assessed on specific tasks: for example, the system described in Godsmark and Brown (1999) is evaluated on two musical tasks: recognizing two interweaved melodies and separating polyphonic music into its constituent melodic lines. An alternative to the SNR improvement is the target-to-interferer ratio (TIR), which is often used to evaluate usable speech measures (Yantorno et al., 2003). User listening tests are also used to assess the quality of the resynthesized/segregated speech signal (Masuda-Katsuse and Kawahara, 1999; Kristjansson et al., 2004). Finally, several algorithms described in this review are only evaluated qualitatively on a small set of examples (Reyes-Gomez et al., 2004; Bach and Jordan, 2005): these approaches are indeed quite advanced and ambitious, and require more time to develop fair evaluation procedures.

Very few experimental results obtained by combining a CASA system with a missing data speech recognizer have been published so far. The main reason may be that it is not the objective of CASA research, or it may be due to the difficulty obtaining good recognition results with such a straight combination. One remarkable paper from this point of view is (Brown et al., 2001), which combines Wang’s neural CASA system with a missing data recognizer, and evaluates it on a noisy version of the TiDigits database. The combined system outperforms spectral subtraction in factory noise, and gives about a 47% recognition rate at 0 dB.

The system described in Barker et al. (2005) is also evaluated on the noisy TiDigits database and gives comparable results to (Brown et al., 2001) at 0 dB, but without any degradation in clean speech. Only a very simple primitive grouping based on an estimate of the local SNR is used in this experiment. Most of the other missing data recognition systems are also evaluated on their recognition accuracy, such as (Kim et al., 2005; Seltzer, 2000; Raj, 2000; Barker et al., 2000; Renevey, 2001). However, the precise comparison of these works is difficult, because of the different experimental setups used. Although the use of the standard Aurora2 database, such as in Morris et al. (2001) gives the opportunity to compare the proposed missing data algorithms with other state-of-the-art noise robust methods, assessing the performances of mask estimation methods in highly non-stationary noise such as background speech from one speaker, or music, cannot be realized with this database.

The cortronic network (Sagi et al., 2001) is also evaluated by computing word recognition rate, in a dedicated experimental setup that merges the speech of 1–20 speakers simultaneously, with approximately the
same energy. The vocabulary is composed of 1024 words. With fewer than five speakers, the recognition rate is above 98%. It decreases to 70% for 10 speakers and down to 20% for 20 speakers. The specific setup of these experiments makes them difficult to compare with other speech recognizers. However, the results look very interesting, especially when considering the novelty of the approach compared to 20 year-old HMM systems.

In this section, we have briefly summarized the main evaluation criteria used in most of the approaches reviewed in this paper. The diversity of these strategies may result from several causes:

- Some of these works, such as CASA or source-separation systems, have not been primarily designed for speech recognition. Nevertheless, they compute masks that can be used in missing data recognition systems.
- It is very difficult and time consuming to build and evaluate a whole speech recognition system, including the mask estimation procedure. When this procedure is independent of the recognizer, it is tempting (and reasonable) to first validate the mask estimation part independently from the recognizer before integrating all the modules.
- Interpreting the global performances may be challenging, because several sources of error contribute to the final word error rate. It is easier to isolate the different parts in order to analyse the weaknesses of the proposed algorithms.

Missing data recognition is a multi-disciplinary and ongoing research area, which partly explains why a clear evaluation procedure has not yet emerged. The best proposal for the time being is to use word recognition accuracy, but this might change in the near future, when speech recognition will be integrated within dialogue and understanding platforms. Higher-level evaluation criteria shall then be considered, such as the Concept Error Rate. For now, a reasonable solution might be to use different evaluation strategies at the different phases of development of a new missing data approach: from qualitative assessment on a few examples to first validate the method, up to standard evaluation procedures when it is fully integrated into the final application.

5.2. Research prospects on missing data masks

CASA has long been referenced by researchers as an interesting front-end to compute masks for missing data speech recognition. However, two factors may question this use of CASA in missing data recognition: first, the recognition rates that have been obtained and published so far with such a combination are not really satisfactory. This may result from the difficulty building efficient CASA systems, or from the challenging task of extracting masks that are useful for the speech recognizer; second, Section 4 presents a range of novel ideas that exploit a statistical and signal-processing background. These ideas represent serious alternatives to CASA in the near future for computing masks for missing data recognition. From this point of view, training stochastic models of masks seem a promising approach to infer efficient masks. However, only basic modelling techniques have been proposed so far. More advanced strategies to model masks are required to improve the performances. This implies for example developing specialized masks for different environments, speaker gender, or clusters of speakers. This would enable the deployment of different strategies, depending on the type of corrupting noise.

Another interesting development would be to reconsider the definition of missing data masks. Different types of masks can be defined and used. For example, masks based on an estimate of local SNRs could be used in a more appropriate way than the classical bounded marginalization. Local SNRs give important information about the marginalization or imputation range, which should be correlated to the estimated SNR and its confidence interval. Furthermore, the optimal masks do not only depend on the signal and the SNR, but should also depend on the acoustic models and the decoder, as shown in Section 2. Exploiting this dependence would lead to new definitions for masks and new methods to compute them.

6. Conclusion

This paper has presented a comprehensive overview of the main approaches to finding spectro-temporal regions dominated by noise, with the objective of applying missing data recognition to recognize the speech
of the main speaker, and with the constraint of using a single microphone. We propose decomposing these approaches in two classes: the first class is related to Computational Auditory Scene Analysis, and is largely inspired by the human auditory system, while the second exploits low-level signal properties and statistical models.

We have shown that the challenge of mask estimation is closely related to several research areas (Computational Auditory Scene Analysis, co-channel speech separation, usable speech, blind source separation and missing data recognition) that make use of similar features, like harmonicity, but consider them from a different perspective. To improve missing data speech recognition, it might be worthwhile to give a global overview of all these methods in order to better understand their similarities and discrepancies. Very few of the reviewed works validate their approaches in terms of recognition accuracy. The challenge is now to exploit this abundant mass of work to give a clear answer to the still unsolved issue of building a baseline missing data recognition system based on effective and robust mask estimation.

The first conclusion one can draw from this review is that both bottom-up and top-down processing are required to estimate masks. This has already been suggested by Bregman, and was proven by a number of experiments depicted in this review. Another conclusion is that most features that can be used for primitive grouping are now clearly identified. A lot of work still has to be done regarding the robust and accurate computation of these features, but it is unlikely that new ones will appear in the near future. It has also been shown that taken individually, each feature is not efficient enough to accurately identify the masks in most common situations; it is therefore very important to exploit several, if not all of these features. Regarding primitive grouping, the efforts should focus now on the computation and combination of features. Besides, there is still a lot of exploratory research to be done concerning the different types of knowledge for schema-driven processing. Stochastic speech models are the most straightforward candidates, but they are not designed for that task, and dedicated models could be more effective. Finally, the combination of primitive features and knowledge models is still an open research problem with great potential for improvement in state-of-the-art missing data recognition solutions.

References


