# Monaural speech separation and recognition challenge

Martin Cooke, John R. Hershey, Steven J. Rennie

▶ **To cite this version:**

Martin Cooke, John R. Hershey, Steven J. Rennie. Monaural speech separation and recognition challenge. Computer Speech and Language, 2009, 24 (1), pp.1. 10.1016/j.csl.2009.02.006 . hal-00598185

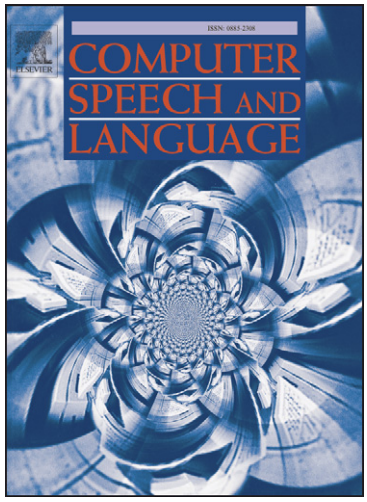
**HAL Id: hal-00598185**
**https://hal.science/hal-00598185**

Submitted on 5 Jun 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript




# Monaural speech separation and recognition challenge

Martin Cooke, John R. Hershey, Steven J. Rennie

PII: S0885-2308(09)00020-5
DOI: 10.1016/j.csl.2009.02.006
Reference: YCSLA 407

To appear in: *Computer Speech and Language*

Received Date: 11 September 2008
Revised Date: 16 February 2009
Accepted Date: 19 February 2009

Please cite this article as: Cooke, M., Hershey, J.R., Rennie, S.J., Monaural speech separation and recognition challenge, *Computer Speech and Language* (2009), doi: 10.1016/j.csl.2009.02.006

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

# Monaural speech separation and recognition challenge

Martin Cooke [a,b,*], John R. Hershey [c], and Steven J. Rennie [c]

[a] *Ikerbasque (Basque Science Foundation)*

[b] *Departamento de Electricidad y Electrónica, Facultad de Ciencias y Tecnología, Universidad del País Vasco, 48940 Leioa, Spain*

[c] *IBM, T. J. Watson Research Center, Yorktown Heights, N.Y., USA*

**Abstract**

Robust speech recognition in everyday conditions requires the solution to a number of challenging problems, not least the ability to handle multiple sound sources. The specific case of speech recognition in the presence of a competing talker has been studied for several decades, resulting in a number of quite distinct algorithmic solutions whose focus ranges from modeling both target and competing speech to speech separation using auditory grouping principles. The purpose of the monaural speech separation and recognition challenge was to permit a large-scale comparison of techniques for the competing talker problem. The task was to identify keywords in sentences spoken by a target talker when mixed into a single channel with a background talker speaking similar sentences. Ten independent sets of results were contributed, alongside a baseline recognition system. Performance was evaluated using common training and test data and common metrics. Listeners' performance in the same task was also measured. This paper describes the challenge problem, compares the performance of the contributed algorithms, and discusses the factors which distinguish the systems. One highlight of the comparison was the finding that several systems achieved near-human performance in some conditions, and one out-performed listeners overall.

*Key words:* speech recognition, speech separation, speaker identification, simultaneous speech, auditory scene analysis, noise robustness.

* Corresponding author

*Email addresses:* m.cooke@ikerbasque.org (Martin Cooke), jrhershe@us.ibm.com (John R. Hershey), sjrennie@us.ibm.com (Steven J. Rennie).

ACCEPTED MANUSCRIPT

## 1 Introduction

Speech recognition by machines in noisy conditions remains an important open problem. Human listeners possess great flexibility in responding to a wide range of acoustic backgrounds and distortions encountered in everyday situations (Assmann and Summerfield, 2004). From the presence of a relatively stationary noise source such as a car engine to the highly time-varying intrusions produced by one or more background talkers, listeners are generally able to communicate successfully. Understanding the basis for robust human performance is important for technological advances in areas such as automatic speech recognition (Gong, 1995) and cochlear speech processors (Dau et al., 2008).

It is well known that human listeners tend to perceive separated 'objects' when listening to an acoustic mixture of sources (Bregman, 1990; Darwin and Carlyon, 1995; Divenyi, 2004). This motivates a source-separation approach for machine listeners. The problem of separating and recognising speech is particularly interesting because it admits a wide variety of very different solution techniques including those based on source independence (e.g. Comon, 1994; Bell and Sejnowski, 1995; Parra and Spence, 2000; Makino et al., 2007), computational auditory scene analysis (e.g. Weintraub, 1986; Ellis, 1996; Wang and Brown, 2006), signal-processing-based speech enhancement (e.g. Benesty et al., 2005; Loizou, 2007) and model-combination (e.g. Varga and Moore, 1990; Gales and Young, 1996). However, until recently, no large-scale global comparison had been undertaken, making it difficult to judge which approaches are worth pursuing and why.

To address this issue, a coordinated speech separation and recognition 'challenge' was initiated in 2006 (Cooke and Lee, 2006). The task for this first challenge was chosen to be non-trivial but at the same time feasible for a range of algorithmic approaches. The goal was to recognise certain keywords in simple sentences produced by one talker when mixed into a single channel with another sentence containing similar material. This problem is not a "real-world" task, in which, for instance, speech is modified in the presence of noise, spoken communication is carried out against a realistic background, rather than one consisting of a single talker emitting highly-similar material, and listeners use binaural cues to help separate and identify the speech of the interlocutor. However, the problem for this first challenge attempt was deliberately constrained in order to provide a best-case scenario for an already difficult problem, since a failure to produce good performance on the constrained task would almost certainly lead to severe problems on more realistic ones. The constrained task also allowed for direct comparison of algorithms and listener performance.

ACCEPTED MANUSCRIPT

The first outputs of the challenge were disseminated at a special session of Interspeech'06 which took place in Pittsburgh, USA. Nine groups presented results on the task (Barker et al., 2006; Deshmukh and Espy-Wilson, 2006; Every and Jackson, 2006; Han et al., 2006; Kristjansson et al., 2006; Ming et al., 2006; Virtanen, 2006; Srinivasan et al., 2006; Schmidt and Olsson, 2006). Revised and extended versions of some of these, together with new contributions, make up the current issue.

Section 2 describes the challenge task, the speech materials distributed to participants, and the scoring procedure. Results obtained by listeners on the same problem are reviewed in section 3, while section 4 describes the performance of a baseline recogniser. The algorithmic approaches adopted by participants are outlined and their results compared in section 5, followed by a discussion of the salient differences between the techniques.

## 2 The challenge problem

### *2.1 Task*

The challenge problem was to recognise keywords from simple *target* sentences when presented with a simultaneous *masker* sentence having a very similar structure. Sentences were recorded separately and mixed additively into a single channel at a range of target-to-masker ratios (TMRs) ranging from +6 dB to -9 dB.

Although the task is not particularly representative of the problems faced in everyday speech perception, it was chosen for the challenge for several reasons. The identification of speech in two-talker conditions from a single microphone is a difficult problem for both listeners and algorithms. While progress has been made on speech recognition in stationary noise, competing speech material provides a highly non-stationary and linguistically-confusing background. In this regard, it is interesting to note that human listeners much prefer non-stationary backgrounds, performing at a significantly lower level when the background is stationary (Bronkhorst and Plomp, 1992; Simpson and Cooke, 2005). Second, behavioural data for this task was already available (Cooke et al., 2008), enabling human-algorithm comparisons. Further, the task focuses on the lower levels of speech processing, making few demands on higher-level linguistic information since all utterances are syntactically, semantically and pragmatically equal, removing one source of variation from human-machine comparisons. Finally, the task employs a simple sentence structure and a relatively small but easily confusable vocabulary, removing the requirement for a large-scale automatic speech recognition infrastructure, and enabling a wider

ACCEPTED MANUSCRIPT

range of participants to contribute algorithms which focus on source separation rather than recognition.

### *2.2 Speech material*

All test and training material was drawn from the GRID corpus (Cooke et al., 2006). GRID sentences conform to the syntax

```
($command $colour $prep $letter $number $coda)
```

with word alternatives given by

```
$command = bin | lay | place | set;
$colour = blue | green | red | white;
$prep = at | by | in | with;
$letter = A | B | C | ... | U | V | X | Y | Z;
$number = zero | one | two ... seven | eight | nine;
$coda = again | now | please | soon;
```

The letter W was not used since it is the only multi-syllabic spoken letter of English

The 34 talkers (18 male, 16 female) in the GRID corpus produced a different set of 1000 sentences each, leading to a total of 34000 unique utterances. All speech material was sampled at 25 kHz and endpointed to remove leading and trailing silence.

A 17000 sentence training set was created by random selection of 500 utterances from each of the 34 talkers. Test sets were drawn from the remaining 17000 utterances. Thus, the test and training set contained the same closed set of 34 talkers.

Seven test sets were created, one clean, the rest composed of sentence pairs added at 6 TMRs (+6, +3, 0, -3, -6, -9 dB). Sentences were paired to be approximately equal in duration. The clean set was produced to enable participants to report baseline recognition performance if they were using something other than the default recogniser provided by the organisers (see section 4). Each of the 6 two-talker test sets contained 600 utterances pairs. Within each test set, roughly equal numbers of utterance pairs came from speakers of different genders (DG), the same gender but different individuals (SG) and the same talker (ST). These TMRs and speaker-pairing conditions were chosen on the basis of earlier two-talker experiments with listeners (Brungart et al., 2001) which demonstrated interesting non-monotonic identification rates as a function of TMR due to the differing availability of cues for speech segregation.

ACCEPTED MANUSCRIPT

In each sentence pair, the color word of the target utterance was always 'white', while the color of the masking utterance was never 'white'. This allowed 'white' to act as a label to distinguish target and masker for listeners and algorithms. The task was to report both the letter and digit keywords spoken by the target talker. The target letter/digit pair was always different from the masker letter/digit pair, but other words in the sentence, such as the coda, were allowed to coincide.

In the distributed test sets, filenames indicated the words spoken by target and masker talkers as well as target and masker talker identifiers. For example, the filename 't4_bwba3p_m9_sgbl4p' denotes the target sentence 'bin white by A 3 please' spoken by talker 4 mixed with the sentence 'set green by L 4 please' spoken by the masking talker, talker 9. Test sets were distributed grouped into directories by TMR. Figure 1 shows auditory spectrograms for this example, mixed at a TMR of 0 dB.

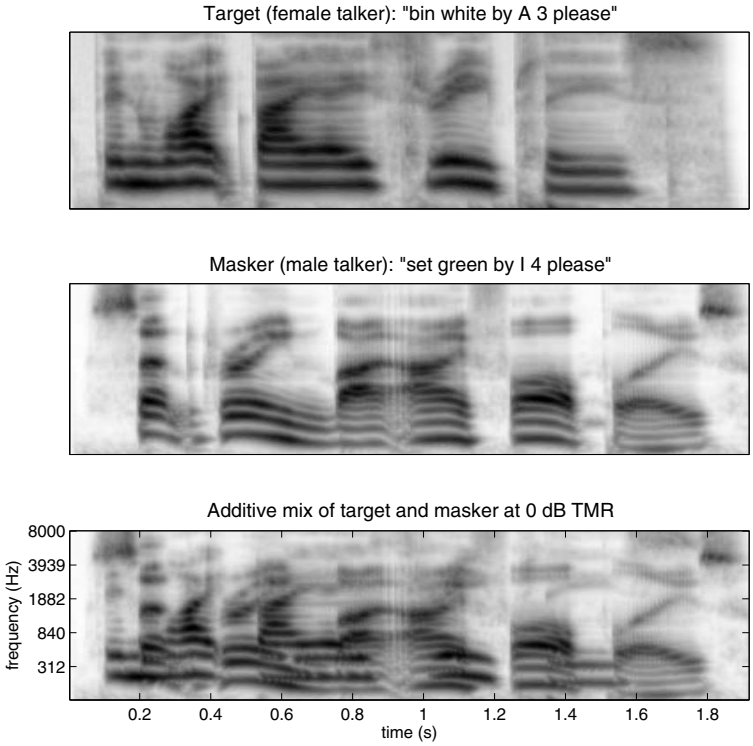



Fig. 1. *Auditory spectrograms for target (top), masker (middle) and their mix (lower) at 0 dB target-to-masker ratio. Auditory spectrograms were produced by filtering the signal using a bank of 200 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz, then sampling the low-pass filtered Hilbert envelope at the output of each filter at 5 ms intervals.*

The clean test set consisted of the target utterances used in the two-talker conditions. To create the two-talker mixtures, the level of the masker was varied to produce the desired token-wise TMR. Consequently, it was possible for participants to recover the clean masker signal simply by subtracting the

ACCEPTED MANUSCRIPT

clean signal from the mixture. This gave participants access to undistorted target and masker signals. Of course, participants were not allowed to use the clean signals or prior information about talker identities or TMRs when reporting their main results. However, participants could use this information to better understand the behaviour of their techniques in various idealised conditions. For example, some participants reported auxilliary results based on having exact prior knowledge of talker identities in order to assess the consequences of errors in talker identification.

Participants were allowed to make use of the constraint that the talkers in the training set were the same as those in the test mixtures, and that they belonged to a closed set.

### *2.3 Scoring procedure*

For each two-sentence mixture, participants' algorithms estimated the most likely letter and digit spoken by the talker who uttered the keyword 'white'. Each mixture was then given a score of 0, 1 or 2 keywords correct. Participants reported the overall percentage keywords correct in each condition. In addition, they reported percentage correct scores for the three subconditions of same talker, same gender, and different gender. A script was provided to generate these measures. In addition, some participants reported speaker identification scores.

## 3 Listener results

As part of a separate study on native versus non-native sentence separation (Cooke et al., 2008), behavioural data for the conditions of the challenge were collected. Eighteen native English listeners with normal hearing identified the letter and digit for the target talker in a block of 100 mixtures for each TMR. Listeners were tested individually in an IAC single-walled acoustically-isolated booth using Sennheiser HD250 headphones at the University of Sheffield. Stimulus presentation and response collection was under computer control. Listeners were familiarized with the stimuli and the task by identifying an independent practice set of 60 sentences in quiet prior to the main set. Condition ordering was randomised across listeners and within each block the 'same talker', 'same gender' and 'different gender' utterance pairings were mixed. To prevent listeners from using absolute level as a cue to the target utterance, presentation level was randomly roved over a 9.5 dB range from stimulus to stimulus. Note that test sets for the algorithmic approaches did not include level variation, and several contributed systems may have benefitted from the

ACCEPTED MANUSCRIPT

constant level of the target utterance (see section 6.7).

Listener results are shown in figure 2 as a function of TMR. The pattern of results is comparable to that found by Brungart et al. (2001) who used sentence pairs with a similar structure and in the same talker-pairing conditions. Listeners had least difficulty in identifying target keywords when the masking talker was of a different gender, and had most difficulty when the same talker was used for target and masker. The strongly nonmonotonic pattern as a function of TMR in the 'same talker' and 'same gender' conditions was also found by Brungart, and is thought to demonstrate the beneficial effect of level differences between target and masker in helping to assign keywords to the target speaker. In the 'different gender' case, other cues are sufficiently strong to render level differences unnecessary.

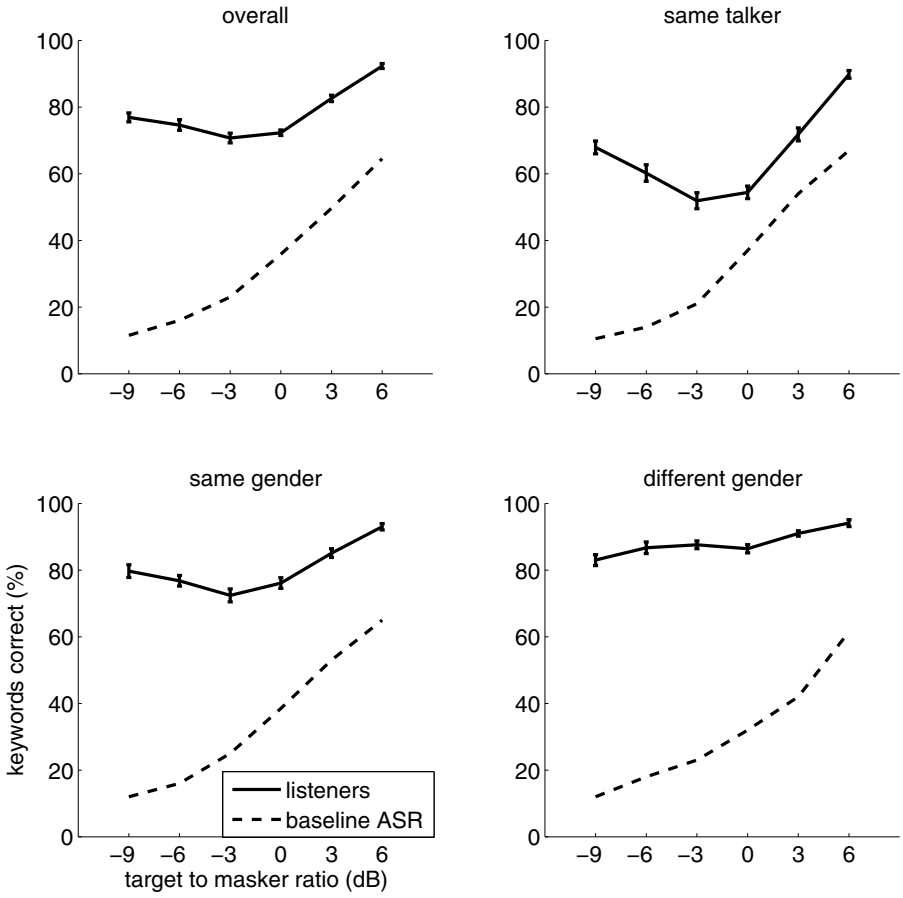



Fig. 2. *Listener and baseline recogniser results on the two-talker datasets. For listeners, error bars indicate ± 1 standard error.*

In general, listeners presumably make use of differences in the talkers' vocal characteristics, grammar, and amplitude to infer the association between the keyword and the letter and digit. However in the same talker condition, with near-zero TMR, these cues are not present. In addition, in the challange task grammar, the distribution of the words in a given position, does not depend

ACCEPTED MANUSCRIPT

on the previous words. This helps explain why listener scores dropped to 52% at -3 dB in the same talker condition, despite the fact that listeners generally have little problem hearing keywords from both target *and* masker sentences (Brungart et al., 2006; Cooke et al., 2008). This score is not much above what would be expected if listeners recognize the words of both talkers correctly, and randomly guess which ones belong to the target utterance.

## 4 Baseline and reference recognition system

A baseline/reference speech recogniser was constructed for the task. The recogniser served a number of purposes. When fed with utterance pairs, it enabled the estimation of lower-bound performance for the task. In addition, it provided those participants who were unable to develop their own recognition architecture with a default recogniser. Finally, those participants whose algorithms led to an output in the form of a time-domain, enhanced signal, were able to use the recogniser directly to evaluate the quality of the estimated speech signal. In fact, the entire process from a set of enhanced waveforms to overall scores in each condition was automated, allowing groups to focus on signal separation rather than speech recognition. Use of the baseline/reference recogniser was optional since some algorithms performed separation and recognition in the same process.

The baseline recogniser was constructed using HTK version 3.1 (Odell et al., 1995). Waveforms were parameterised into standard 39-dimensional *Mel frequency cepstral coefficients* (MFCCs), i.e. 12 Mel-cepstral coefficients and the logarithmic frame energy plus the corresponding delta and acceleration coefficients (MFCC_E_D_A). The 51 words required to support GRID utterances were modelled as whole-word *hidden Markov models* (HMMs) with a left-to-right model topology, with no skips over states and 32 Gaussian mixtures per state and diagonal covariance matrices. The number of states for each word was based on 2 states per phoneme: the spoken letters, 'at', 'by' and 'in' used 4-state models while all remaining words employed 6-state models apart from 'again' and 'zero' (8 states) and 'seven' (10 states).

Baseline recognition results are shown in figure 2. Listener performance far exceeds that of the baseline recogniser in most conditions by a large margin. For example, in the different gender condition at a TMR of -9 dB, the difference is 71 percentage points. This is not surprising since the recogniser has no basis on which to select the target utterance. It is interesting to note, however, that the difference reaches its minimum (17 percentage points) in the same talker condition at 0 dB and 3 dB. The baseline recogniser performance at these TMRs is well above chance. This is a clear indication that fragments of letter and digit words from the masker and target utterances were sufficiently

ACCEPTED MANUSCRIPT

'audible' to allow a recogniser which is unable to perform any kind of speech separation to nevertheless identify correctly one or other keyword on most occasions.

The pattern of performance with TMR was quite different for the listeners and baseline recognition system. While listeners displayed a non-monotonic pattern, the recogniser suffered a continual degradation as the masker level increases relative to the target. As mentioned earlier, listeners can exploit a level difference cue to improve performance at negative TMRs in spite of increasing amounts of energetic masking, while the recogniser is most likely to respond to the energetically-dominant utterance, a 'strategy' which helps for positive TMRs but causes errors at negative TMRs. A further listener-baseline difference is revealed by the ranking of the three speaker pairing subconditions. Listeners performed best in the different gender case and worst when the same talker was used for each sentence in the pair. The baseline recogniser worked best for the same talker condition for TMRs 0-6 dB and performed least well in the different gender condition at these TMRs. This pattern was reversed for the negative TMRs.

## 5 Results

Table 1 summarizes the overall performance of the systems in this issue, along with some others that addressed the same problem, but were published elsewhere. The accuracy of human listeners on the task is also reported, as is the basic approach of each system. For systems published in this issue, the performance on each of the sub-tasks is plotted in Figure 3 as a function of target to masker ratio.

Overall, all systems described in the current issue performed significantly above the baseline. However, there was a great deal in variation in performance across systems, ranging from 20 percentage points at the most favourable TMR to around 50 percentage points in the least favourable condition. Three systems (Weiss and Ellis, 2009; Li et al., 2009; Shao et al., 2009) fell into a relatively narrow performance range but three others scored at a higher level. Notably, the system of Hershey et al. (2009) outperformed all others across the range of TMRs, and beat listeners at intermediate TMRs. This advantage over listeners was largely based on a significantly higher score in the same gender condition. In a subset of (condition, TMR) pairings, the performance of the Barker et al. (2009) system was indistinguishable from listeners. At the extreme TMRs where the target speaker was at its most or least dominant, listeners outperformed all contributed algorithms overall. When compared against the best system, the listener advantage was largely due to superior performance in the same talker condition.

ACCEPTED MANUSCRIPT

| System | Approach | Accuracy |
|---|---|---|
| Hershey et al. (2009) | Model-based, joint decoding | 78.4% |
| Human Listeners | Listening | 77.7% |
| Virtanen (2006) † | Model-based, alternating decoding | 65.8% |
| Barker et al. (2009) | CASA, missing features | 63.8% |
| Ming et al. (2009) | Model-based, missing features | 58.4% |
| Schmidt et al. (2006) † | Non-neg. matrix factorization | 50.2% |
| Weiss and Ellis (2009) | Model-based, joint decoding | 48.0% |
| Li et al. (2009) | Model-based, reconstruction | 47.7% |
| Shao et al. (2009) | CASA, missing features | 45.5% |
| Baseline Recognizer | HTK recognizer, no enhancement | 33.4% |
| Deshmukh et al. (2006) † | Phase opponency enhancement | 31.6% |
| Every et al. (2006) † | Pitch-based enhancement | 23.3% |
| Chance | Guessing | 7.0% |

Table 1
Overall word error rate (%) of the ASR systems that were entered into the Pascal 2006 Speech Separation Challenge. † not published in this issue.

The characteristic non-monotonic shape of the performance-TMR function for listeners was not echoed at the level of overall scores by any of the algorithms. However, two systems (Hershey et al., 2009; Barker et al., 2009) demonstrated similar non-monotonicity in the same talker condition, suggesting that they were able to use level differences between target and masking talkers to improve performance even when the target talker was the weaker of the two.

## 6 Discussion

The monaural speech separation and recognition problem has been approached in these works in a variety of ways. At the broadest level, contributed systems differed according to the assumptions and constraints of the task that were taken into account. In general one can expect a trade-off between the performance on the challenge task, and the applicability of a system to other less constrained scenarios as a function of the number of task constraints that were used. Systems ranged from those that took advantage of few constraints other than the fact that the target is a human voice, to those that took advantage of almost every available constraint, from the closed speaker set to the known

ACCEPTED MANUSCRIPT

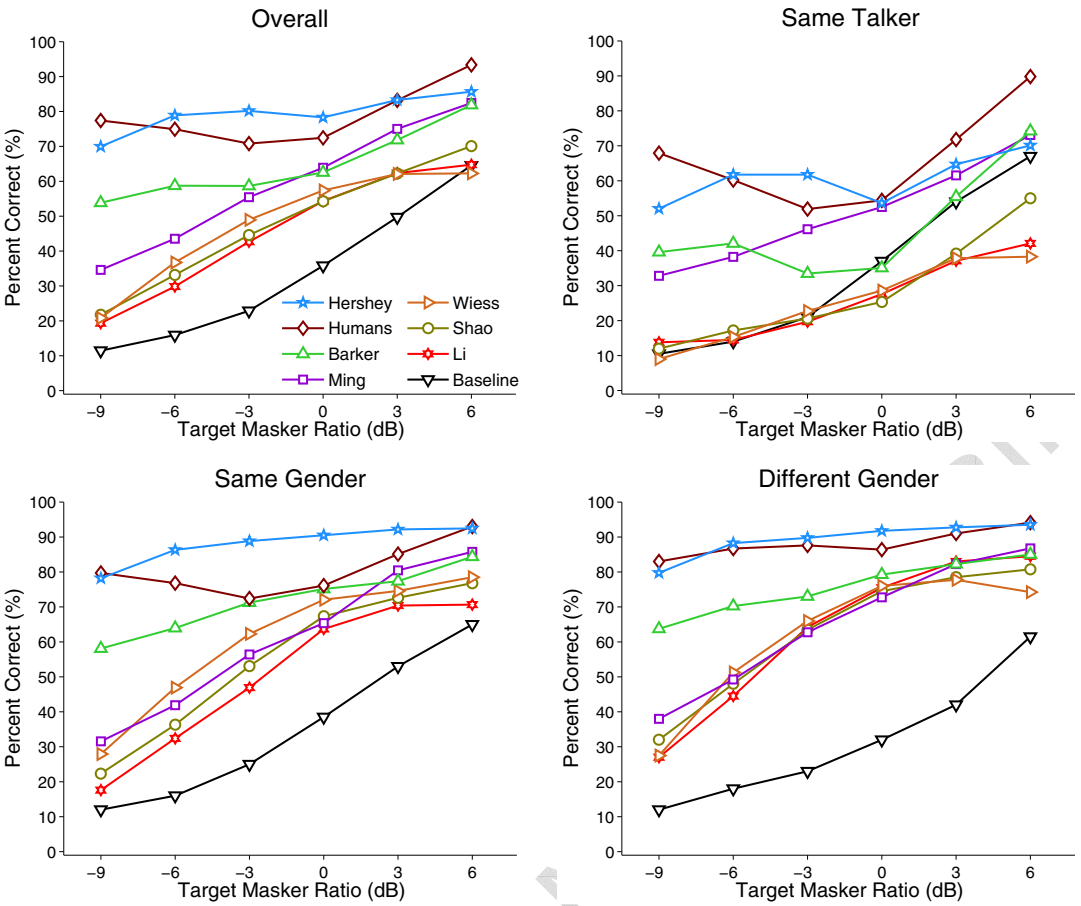



Fig. 3. *Recognition results on the two-talker data sets for all systems published in this issue.*

task grammar for both speakers. Thus it is useful to look at the approaches in terms of the assumptions and strategies they used.

### *6.1 Model-based versus CASA approaches*

The most significant difference in approach is between the model-based and the *computational auditory scene analysis* (CASA) approaches. In the model-based approaches, such as (Hershey et al., 2009; Virtanen, 2006; Li et al., 2009; Ming et al., 2009; Weiss and Ellis, 2009), *top-down* generative models are used to capture the statistics of features of isolated signals, as well as the effect on the features of combining two signals. Inference then seeks the two speech signals that are most likely given the observed mixture. The decomposition of the spectrogram (or other time-frequency representation) into its constituent sources emerges as a by-product of this inference.

In CASA, (Barker et al., 2009; Shao et al., 2009) the approach is *bottom-up*: segmentation and grouping rules operate on low-level features to determine

ACCEPTED MANUSCRIPT

which regions of a spectrogram belong to the same source. Grouping cues include common pitch, common onset, common amplitude modulation, temporal continuity, and so on (Wang and Brown, 2006). Rather than segmenting the two sources completely, however, the signal is typically divided into local regions, and then passed on to a model-based system which disambiguates the global segmentation. This involves an assumption that there are relatively large regions of the spectrogram that are dominated by one or another source. In contrast, model-based approaches typically do not make any assumptions about the regions over which one signal may dominate another, other than implicitly by choosing a spectrogram with a particular frequency or time resolution. Many model-based approaches, moreover, explicitly model how the sources interact to better represent and separate mixed features that are not dominated by a single source, as described in section 6.3 below.

The CASA approach has the advantage that it may generalize to many different types of signals. However, it also has the disadvantage that the grouping rules are not usually trained on data (see Bach and Jordan, 2006, for work in this direction). In cases where much is known about the signals, such as in the speech separation scenario, it is easier to incorporate that information into a model-based framework than in a CASA framework.

### *6.2 Speech separation versus mask inference*

Another important distinction is between systems that reconstruct the speech signals or features for use in a standard recognizer, and those that estimate the reliability of different areas of the original spectrogram for use in a specialized *missing-feature* recognizer (Cooke et al., 2001). A related approach, known as *uncertainty decoding*, involves reconstructing speech features along with an estimate of their uncertainty, and taking this into account when doing recognition (Droppo et al., 2002). All of these methods suffer from the mismatch between the feature domains that are useful for speech separation and those that are useful for isolated speech recognition, although each has a different solution to this problem.

In reconstruction-based methods the audio signal of each speaker is estimated and then passed to a recognizer (Hershey et al., 2009; Virtanen, 2006; Weiss and Ellis, 2009; Li et al., 2009; Ming et al., 2009). In these methods, different models can be used for separation and recognition. Speech recognition systems typically use full language models or grammars, model multiple frames at a time, and model in a transformed domain (MFCCs) that corresponds to using full covariances in the spectral domain. The full covariances in the spectral domain are particularly difficult to deal with since signal combination models operate in the spectral domain. Instead, separation systems often make use of

ACCEPTED MANUSCRIPT

single-frame models with diagonal covariances in the spectrum. In addition, experiments can be done using simplified dynamics for separation, instead of the full grammar used for recognition. The differences in features are important: Hershey et al. (2009) report that even when the separation model has the full task grammar for both speakers, and hence performs recognition in the course of reconstruction, the best result is obtained by using it for reconstruction and passing the result to a single-speaker recognition system. A drawback of reconstruction is that using a point estimate of the signal disregards any information we have about the reliability or uncertainty of different parts of the signal. Note however that in the process of reconstruction, the model-based approaches implicitly consider all masking functions allowed by the model.

Missing feature approaches in general are even more flexible, in the sense that only a masking function need be inferred prior to single-speaker recognition. The recognizer then integrates out features labeled as masked or unreliable. In Barker et al. (2009), rather than passing a single masking function to the recognizer, an initial segmentation is used to produce a set of speech *fragments*. A large number of possible global masking functions can be constructed by combining different subsets of the segmented features (Barker et al., 2005). The recognizer then uses dynamic programming to explore the large space of all possible masking functions (a related approach was used in Reyes-Gómez and Jojic, 2006). The approach of Ming et al. (2009) also uses the recognizer in inferring the masking functions, but does so without an initial segmentation. Instead, it approximately computes the likelihood of the best masking function for each frame by summing the likelihood over sets of masking functions defined by the number of masked features, and choosing the set with the largest sum. One drawback of the missing feature approach is that the features used must have the property that signals overlap as little as possible. In the feature spaces used by most speech recognizers, this property does not hold.

Uncertainty decoding seeks to avoid the problems of both reconstruction and missing feature approaches, by preserving the uncertainty in the reconstruction, and passing this along to the recognizer. In Shao et al. (2009), a masking function is used along with a GMM speech model to reconstruct the masked portions of the signal, and estimate the uncertainty of their reconstruction. These uncertainties are then used to modify the recognition models in the feature domain of the recognizer. A drawback of this approach is that even if the uncertainties in the masking domain have diagonal covariance, in general it is difficult to transform these uncertainties into the feature space of the recognizer without requiring full covariance models. The main problem is that the uncertainty of a single spectral band is spread out in the cepstral domain by the discrete cosine transform, so that without full covariances, local uncertainty becomes global uncertainty.

ACCEPTED MANUSCRIPT

### *6.3 Signal interaction modeling*

In model-based approaches one has to compute the likelihood of the observed signals, given models of the two speakers. For reconstruction-based systems, one also has to compute an estimate of the hidden speaker signals. Because signals are typically modeled in the log power spectrum, the exact signal interaction model is highly nonlinear, and various approximations are necessary. The *Algonquin* method Frey et al. (2001) iteratively linearizes the interaction function, and approximates the posterior of the two signals with a Gaussian for each combination of states of the two speaker models. The *max model* approximates the logarithm of the sum of the two power spectra using the logarithm of the maximum of the power spectra. Hershey et al. (2009) explored both Algonquin and a max interaction model, with Algonquin working slightly better than the max model.

A simpler version of the max model known as the *max vector-quantizer* (MAXVQ, Roweis (2003)) model was used in Li et al. (2009) and Weiss and Ellis (2009). In this version of the max model, a given signal either dominates the other and hence is observed, or otherwise is masked. The observed parts of the signal need not be estimated; for the masked portions, usually the prior mean of the signal for a given state or mixture component is used as an estimate, as in Weiss and Ellis (2009). However, in Li et al. (2009) the signal estimates are taken to be the prior mean of the winning component. Virtanen (2006) in contrast, computed likelihoods using *parallel model combination* (PMC), which uses a moment-matching approximation to the likelihood of the log-sum of two Gaussians. In this case the posterior of the hidden speaker signals for a given state cannot be computed, and instead the prior mean is used as an approximation. An interesting departure from these cases is the non-negative matrix factorization approach of Schmidt and Olsson (2006), which models signals directly in the spectrum as a sum over basis functions. In this case the interaction model is linear in the model domain, and inference is greatly simplified.

In Ming et al. (2009) and Barker et al. (2009), missing feature approaches are used to avoid an explicit interaction model. The *speech fragment decoding* technique used in Barker et al. (2009) assumes, for a given masking pattern, that the masked regions have an energy less than the observed signal, which is consistent with the max model. In contrast, the *sub-band union model*, introduced in Ming and Smith (2003) and employed by Ming et al. (2009), entails a weaker assumption about interaction, in which time-frequency regions that have high probability under the model are considered dominant, and low-probability regions are considered masked, regardless of their relative amplitudes, an assumption also used in Reyes-Gómez and Jojic (2006).

ACCEPTED MANUSCRIPT

In approaches that avoid explicit signal models, such as CASA components, there still may be an implicit signal combination model. For instance in the CASA components of Barker et al. (2009) and Shao et al. (2009) it is assumed that one source dominates in given time-frequency bin, while the other is masked, which again suggests a max model.

### *6.4 Temporal modeling*

A major point of difference between approaches hinges on how the dynamic properties of speech are used. At the highest level, the task defines a word grammar for both the target and masker signals. Words are made of phonemes, which describe the temporal evolution of aspects of the signal associated with the pronunciation. There also may be dynamics associated with *paralinguistic* components of speech; that is, any features of a speech signal that are not represented by the phonetic pronunciation of the words. For instance, the pitch and amplitude of an utterance may follow patterns that are not inherent to the sequence of words being spoken. All of the approaches ultimately use a speech recognizer with the task grammar as the final recognizer. However, in the processing stages leading up to the final recognition, a variety of these dynamical constraints are used in both the CASA and model-based approaches. Which constraints are used, whether or not they are applied to both of the speakers, and how inference is done using the dynamics, tend to determine how accurately, and how efficiently, each system can perform.

The two best performing systems (Hershey et al., 2009; Virtanen, 2006) used the task grammar for both the target and the background signal to form a two-dimensional HMM model of the entire task, with one Markov chain per talker. The main difficulty for such models is the computational complexity of exact inference. In Virtanen (2006) approximate inference was done by alternating between the two Markov chains, holding the state sequence of one constant while computing the Viterbi algorithm on the other. In Hershey et al. (2009) this was compared with other techniques, including the exact 2-D Viterbi algorithm, and an efficient loopy belief propagation algorithm. The exact algorithm worked best, but the loopy belief propagation algorithm was within 3% in terms of absolute accuracy, despite being orders of magnitude more efficient. Hershey et al. (2009) also investigated different levels of dynamics, and showed decreases in performance with less constrained dynamics. One of the experiments relaxed the dynamics of just the background model by using a bag-of-words model for the background, resulting in an increase in overall error rate of 4.2% absolute. In Weiss and Ellis (2009) a similar 2-D HMM system used phoneme-loop dynamics instead of the task grammar, for separation, but used the task grammar to recognize the separated signals.

ACCEPTED MANUSCRIPT

The other approaches have used less tightly coupled models for target and background. In Ming et al. (2009), the sub-band union model, which has a less coupled interaction model than the above methods, only uses task dynamics for one signal at a time. After recognizing the dominant speaker, the result is used to reconstruct an estimate of the background speaker signal, which in turn is recognized using the task grammar. In the missing-feature approach of Barker et al. (2009), the task grammar dynamics determine the final segmentation, but the initial segmentation is done using only low-level dynamics such as pitch contiguity, and voicing state transition.

In contrast to the above, the approach of Shao et al. (2009) uses only low-level grouping cues, such as pitch, to produce an initial segmentation, followed by grouping based on speaker-dependent GMM models. Thus the task grammar is only used in the uncertainty decoding recognizer, after reconstruction has been done. Similarly the model-based approach of Li et al. (2009), uses only single-frame models with no dynamics to reconstruct the separated speech signals.

### *6.5 Speaker-dependent modeling*

The majority of the systems that participated in the challenge used speaker-dependent models of both the target and masking speakers to do separation, taking advantage of the constraint that the test speakers were selected from a known set of 34 speakers, for which isolated training data was provided. Exceptions to this trend included the system presented in Ming et al. (2009), which utilized a speaker-independent model for the weaker source in the mixture, and the system of Barker et al. (2009), which used speaker-dependent models to represent the target source, but does not assume that the masking signal is speech. To utilize speaker-dependent models requires that the speakers present in the mixture be identified in a tractable manner. The speaker identification systems used by challenge contestants are discussed in section 6.6.

While the assumption that the target speaker is *enrolled* (and conforms to a specific grammar) is reasonable in many applications, such assumptions about the masker are less applicable. In Hershey et al. (2009), the authors investigated the effect of relaxing these assumptions about the masker on system performance. When a gender-dependent background acoustic model with just 256 components was used, the error rate increased by 4.1%. When a speaker-independent background model with just 256 components was used, the error rate increased by 13.8%.

In many application domains, only a few seconds of data are provided during

ACCEPTED MANUSCRIPT

speaker enrollment, or it cannot be assumed that the target speaker has enrolled. Adaptation is a natural way to try and generalize and therefore improve the performance of any model-based system. In speech recognition, model adaptation is a heavily studied research area, and is known to be critical to achieving state-of-the-art performance (Leggetter and Woodland, 1994; Gales, 1998; Rennie et al., 2006). In Weiss and Ellis (2009) the authors demonstrated that for out-of-speaker-set test data, using the best speaker-dependent model from the training set does not work as well as adapting a linear combination of the speaker-dependent models to best fit each test utterance. Continued research into the generalization of the systems that participated in the speech separation challenge to very large or open speaker sets is an important direction of future investigation.

### *6.6 Speaker identification*

In speaker-dependent model-based systems it is necessary to determine the set of speakers that are present in the utterance. An exhaustive search for the best speaker combination scales exponentially in number of speakers in the mixture, and therefore is prohibitive to compute, even for two speakers. In the challenge, there are 2 speakers in each mixture and 34 speakers to choose from for each speaker, and therefore $34^2 = 1156$ possible speaker combinations to consider.

In Hershey et al. (2009), speakers were identified by using a single bank of speaker-dependent gaussian mixture models to represent mixed speech in the high resolution log frequency domain. Under the model it is assumed that each frame is either dominated by a single source, or is unreliable. This approach avoids considering combinations of speaker models, and exploits the sparsity of the speech signal over time to identify what speakers are present. Frames that lead to posterior speaker distributions that are not discriminative are assumed to be unreliable and are discarded. An estimate of the posterior distribution of the speaker identities for each utterance is obtained by averaging over the posteriors of reliable frames. This posterior estimate is then used to select a small subset of likely speaker combinations to evaluate with a more accurate speaker interaction model. The baseline system Hershey et al. (2009) predicted both the target and masker speakers correctly in greater than 98% of all test utterances. Virtually the same system was adopted in Weiss and Ellis (2009) and Li et al. (2009). In Li et al. (2009) the system was further tuned, yielding an accuracy of greater than 99%.

Barker et al. (2009) investigated a more sophisticated speaker identification algorithm that exploits the sparsity of speech over time and frequency and integrates the task dynamics into the estimation. The system is based upon

ACCEPTED MANUSCRIPT

a fragment-decoder with a CASA front-end, which selects what fragments are reliable based on the constraints of the target speech model. Only the target source is explicitly modeled, and therefore only the target speaker needs to be inferred, and so searching over combinations of speaker identities is not necessary. By selecting the speaker that has the highest *rate of increase* in likelihood in the target-identifying part of the grammar ('white'), rather than selecting the speaker with the highest likelihood on the data, the target speaker identification accuracy of their system improved from 88.9% to 93.7%, and the word accuracy of their system improved from 59.1% to 63.8%. In Ming et al. (2009), the dominant speaker was identified by doing missing-feature Viterbi decoding on the input mixture, using a single bank of bi-gram HMM speaker models. Here again, taking a missing-feature approach to inference avoids having to consider combinations of speakers during inference.

An alternative approach is to iteratively estimate the identities of the speakers to avoid considering all possible combinations of speaker models. This was implicitly done in Virtanen (2006) by using a bank of speaker HMMs for both the target and masking sources coupled by a PMC interaction model, and iteratively decoding one source given the current state-sequence estimate of the other source. Speaker identification accuracies were not reported, but the approach was clearly effective: the performance of the system on the challenge task was exceeded only by Hershey et al. (2009), and human listeners.

In Shao et al. (2009), speaker-dependent GMM models of both the target and masker in the gammatone-filter (GF) domain were used to simultaneously group CASA-derived segments and determine the identities of the speakers, using an approximate beam search over speaker identities and foreground-background assignments. The overall target speaker identification accuracy of the system is 90.6%, and 46.4% for both speakers.

### *6.7 Speaker gain estimation*

The test mixtures in the task were mixed additively at 6 TMRs ranging from -9 dB to 6 dB in 3 dB increments. The gains of the speakers were unknown at test time, and therefore any speaker models that are used during speech separation or recognition must be normalized to match the mixture data.

Ultimately, the efficacy of model-based speech separation and recognition systems hinges upon the accuracy of the speech model(s). When speaker-dependent models are utilized, the accuracy of the speaker identities depends upon the accuracy of their gains, and vice versa, which in turn ultimately determines how representative the models used for speech separation and recognition are.

ACCEPTED MANUSCRIPT

The speaker identification system presented in Hershey et al. (2009) and described in the previous section used a quantized representation of gain to make the system gain-robust, and then jointly estimated the most probable combination of speaker identities and gains on a small set of probable speaker combinations using a more accurate interaction model. This speaker identification system was also used by Weiss and Ellis (2009) and Li et al. (2009). In contrast in Virtanen (2006), the gains of the sources were iteratively estimated as part of the iterative source decoding procedure.

In Barker et al. (2009), knowledge of the fact that the absolute gain of the target speaker is constant over all TMRs in the training *and* test data was used, and the gain of the target speaker was *not* adapted. This undoubtedly makes the recognition problem somewhat easier, and unfortunately it is not clear which other contestants (if any) were aware and knowingly took advantage of this. In Hershey et al. (2009) (and presumably in Weiss and Ellis, 2009; Li et al., 2009, this information was not leveraged).

In Ming et al. (2009), the gain of the dominant source appears to be estimated only after decoding to reconstruct the weaker source. This system therefore probably also benefits from the way that the data was artificially mixed when the TMR is above 0 dB. In Shao et al. (2009), there is no mention of if and how the gains of the speakers were estimated. Poor masker gains may explain the poor background speaker identification results.

### *6.8 Limitations and future challenges*

The current challenge represents a step towards large-scale evaluations in the field of robust speech processing, but has a number of limitations. The task, although challenging for humans and algorithms alike, is not representative of real-world speech understanding in adverse conditions, for a number of reasons. A single competing talker from a known, small, closed set of talkers, while interesting from a behavioural perspective, represents only one small corner of the space of background conditions in which speech is typically processed. Similarly, in practice, the additive combination of target speech and background noise makes the simplifying assumption that the two are independent, while it is known that speakers make quite complex modifications to their productions in the presence of noise (Junqua, 1993). Other limitations of the current task include the small vocabulary and highly-constrained grammar. Finally, the fact that speech and background were mixed into a single channel provided no opportunities for those approaches which exploit two or more sensors, but at the same time obviated the need to handle the difficult problem of reverberation. Ongoing and future challenges will relax some of these limitations. For example, Vincent et al. (2007) describes an evaluation campaign

ACCEPTED MANUSCRIPT

for stereo audio source separation.

## 7 Conclusions

Speech understanding in the presence of an interfering talker is a challenging problem for listeners and algorithms. This article introduced an evaluation task based on identifying keywords in sentences spoken by a target talker when mixed with similar sentences produced by a masking talker. Submitted systems were typically based on models for the target and background speech or were inspired by auditory scene analysis. For the current task, where precisely two speakers from a known, closed set were mixed, model-based approaches performed well, with one such system out-performing human listeners overall. The evaluation task has a number of limitations which include the use of restricted vocabulary, syntax and talkers, single-channel mixing under the assumption of target/masker independence, and a single background condition. Future evaluations will have to relax some of these constraints to approach the scientific goal of speech understanding in everyday adverse conditions.

## Acknowledgements

The authors thank the following the EU Network of Excellence PASCAL (Pattern Analysis, Statistical modeling and Computational Learning) (www.pascal-network.org) for funding to support the Challenge, Ning Ma and Youyi Lu for their help in producing the baseline recogniser and scoring scripts, Jon Barker for constructive suggestions and assistance with stimulus construction, and Peder Olsen for many helpful discussions. Portions of this work were carried out while the first author was at the University of Sheffield.

## References

Assmann, P., Summerfield, Q., 2004. The perception of speech under adverse acoustic conditions. In: Greenberg, S., Ainsworth, W. A., Popper, A. N., Fay, R. R. (Eds.), Speech Processing in the Auditory System. Springer Handbook of Auditory Research Vol. 18.

Bach, F. R., Jordan, M. I., 2006. Learning spectral clustering, with application to speech separation. J. Mach. Learn. Res. 7, 1963–2001.

Barker, J., Coy, A., Ma, N., Cooke, M. P., 2006. Recent advances in speech fragment decoding techniques. In: Proc. Interspeech 2006. Pittsburgh, pp. 85–88.

ACCEPTED MANUSCRIPT

Barker, J., Ma, N., Coy, A., Cooke, M., 2009. Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. Computer Speech and Language (this issue).

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

Bell, A. J., Sejnowski, T. J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7 (6), 1129–1159.

Benesty, J., Makino, S., Chen, J. (Eds.), 2005. Speech Enhancement. Springer.

Bregman, A. S., 1990. Auditory Scene Analysis. MIT Press, Cambridge MA.

Bronkhorst, A. W., Plomp, R., 1992. Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing. Journal of the Acoustical Society of America 92, 3132–3139.

Brungart, D. S., Chang, P. S., Simpson, B. D., Wang, D. L., 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. Journal of the Acoustical Society of America 120, 4007–4018.

Brungart, D. S., Simpson, B. D., Ericson, M. A., Scott, K. R., 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. Journal of the Acoustical Society of America 100, 2527–2538.

Comon, P., 1994. Independent component analysis, a new concept? Signal Processing 36 (3), 287 – 314.

Cooke, M. P., Barker, J., Cunningham, S. P., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America 120, 2421–2424.

Cooke, M. P., Garcia Lecumberri, M. L., Barker, J. P., 2008. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. Journal of the Acoustical Society of America 123, 414–427.

Cooke, M. P., Green, P. D., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and uncertain acoustic data. Speech Communication 34 (3), 267–285.

Cooke, M. P., Lee, T. W., 2006. Speech separation challenge website. `http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm`.

Darwin, C. J., Carlyon, R. P., 1995. Auditory Grouping. In: The Handbook of Perception and Cognition. Vol. 6 Hearing. Academic Press.

Dau, T., Buchholz, J., Harte, J., Christiansen, T., 2008. Auditory signal processing in hearing-impaired listeners. Centertryk.

Deshmukh, O., Espy-Wilson, C., 2006. Modified phase opponency based solution to the speech separation challenge. In: Proc. Interspeech 2006. Pittsburgh, pp. 101–104.

Divenyi, P. (Ed.), 2004. Speech separation by humans and machines. Springer.

Droppo, J., Deng, L., Acero, A., 2002. Uncertainty decoding with splice for noise robust speech recognition. IEEE Conf. Acoust. Speech Signal Processing.

Ellis, D. P. W., 1996. Prediction-driven computational auditory scene analysis.

ACCEPTED MANUSCRIPT

Ph.D. thesis, MIT, Cambridge MA.

Every, M. R., Jackson, P. J. B., 2006. Enhancement of harmonic content of speech based on a dynamic programming pitch tracking algorithm. In: Proc. Interspeech 2006. Pittsburgh.

Frey, B., Deng, L., Acero, A., Kristjansson, T., September 2001. Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In: Eurospeech.

Gales, M., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. CSL 12.

Gales, M., Young, S., 1996. Robust continuous speech recognition using parallel model combination. IEEE Transactions on Speech and Audio Processing 4, 352–359.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Communication 16 (3), 261–291.

Han, R., Zhao, P., Gao, Q., Zhang, Z., Wu, H., Wu, X., 2006. Casa based speech separation for robust speech recognition. In: Proc. Interspeech 2006. Pittsburgh.

Hershey, J. R., Rennie, S. J., Olsen, P. A., Kristjansson, T. T., 2009. Superhuman multi-talker speech recognition: A graphical model approach. Computer Speech and Language (this issue).

Junqua, J.-C., 1993. The lombard reflex and its role on human listeners and automatic speech recognizers. JASA 93, 510–524.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., 2006. Superhuman multi-talker speech recognition: The IBM 2006 Speech Separation Challenge system. In: Proc. Interspeech 2006. Pittsburgh.

Leggetter, C. J., Woodland, P. C., 1994. Speaker adaptation of continuous density HMMs using linear regression. ICSLP, 451–454.

Li, P., Guan, Y., Wang, S., Xu, B., Liu, W., 2009. Monaural speech separation based on MAXVQ and CASA for robust speech recognition. Computer Speech and Language (this issue).

Loizou, P. C., 2007. Speech Enhancement: Theory and Practice. CRC Press.

Makino, S., Lee, T. W., Sawada, H. (Eds.), 2007. Blind Speech Separation. Springer.

Ming, J., Hazen, T., Glass, J., 2009. Combining missing-feature theory, speech enhancement, and speaker-dependent/independent modeling for speech separation. Computer Speech and Language (this issue).

Ming, J., Hazen, T. J., Glass, J. R., 2006. Combining missing-feature theory, speech enhancement and speaker-dependent/-independent modeling for speech separation. In: Proc. Interspeech 2006. Pittsburgh.

Ming, J., Smith, F. J., 2003. Speech recognition with unknown partial feature corruption  a review of the union model. Computer Speech and Language 17, 287–305.

Odell, J., Ollason, D., Woodland, P., Young, S., Jansen, J., 1995. The HTK Book for HTK V2.0. Cambridge University Press, Cambridge, UK.

Parra, L., Spence, C., May 2000. Convolutive blind source separation of non-

ACCEPTED MANUSCRIPT

stationary sources. IEEE Trans. Speech and Audio Processing, 320–327.
Rennie, S., Kristjansson, T., Olsen, P., Gopinath, R., 2006. Dynamic noise adaptation. ICASSP.
Reyes-Gómez, M. J., Jojic, N., December 2006. Signal separation by efficient combinatorial optimization. In: Advances in Models for Acoustic Processing NIPS 2006 workshop.
Roweis, S., 2003. Factorial models and refiltering for speech separation and denoising. Eurospeech, 1009–1012.
Schmidt, M. N., Olsson, R. K., 2006. Single-channel speech separation using sparse non-negative matrix factorization. In: Proc. Interspeech 2006. Pittsburgh.
Shao, Y., Srinivasan, S., Jin, Z., Wang, D., 2009. A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech and Language (this issue).
Simpson, S. A., Cooke, M. P., 2005. Consonant identification in n-talker babble is a nonmonotonic function of n. Journal of the Acoustical Society of America 118, 2775–2778.
Srinivasan, S., Shao, Y., Zhaozhang, J., Wang, D., 2006. A computational auditory scene analysis system for robust speech recognition. In: Proc. Interspeech 2006. Pittsburgh.
Varga, A. P., Moore, R. K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP 1990. pp. 845–848.
Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J., 2007. First stereo audio source separation evaluation campaign: data, algorithms and results. LNCS 4666, 552–559.
Virtanen, T., 2006. Speech recognition using factorial hidden markov models for separation in the feature space. In: Proc. Interspeech 2006. Pittsburgh.
Wang, D.-L., Brown, G. J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms and Applications. IEEE Press/Wiley-Interscience.
Weintraub, M., 1986. A computational model for separating two simultaneous talkers. In: Proc. ICASSP 1986. pp. 81–84.
Weiss, R. J., Ellis, D., 2009. Speech separation using speaker-adapted eigenvoice speech models. Computer Speech and Language (this issue).

ACCEPTED MANUSCRIPT