Distant Speech Recognition: No Black Boxes Allowed

John McDonough[1]*, Matthias Wölfel*[2]*, Kenichi Kumatani*[1]*, Barbara Rauch*[1]*, Friedrich Faubel*[1]*, and Dietrich Klakow*[1]

[1]Spoken Language Systems, Saarland University, Saarbrücken, Germany
Email: {john.mcdonough,kkumatani,barbara,ffaubel,klakow}@lsv.uni-saarland.de
[2]Institut für Theoretische Informatik, Universität Karlsruhe (TH), Karlsruhe, Germany
Email: wolfel@ira.uka.de

Web: http://distant-speech-recognition.org

Abstract

A complete system for *distant speech recognition* (DSR) typically consists of several distinct components. While it is tempting to isolate and optimize each component individually, experience has proven that such an approach cannot lead to optimal performance. In this talk, we will discuss several examples of the interactions between the individual components of a DSR system. In addition, we will describe the synergies that become possible as soon as each component is no longer treated as a "black box". To wit, instead of treating each component as having solely an input and an output, it is necessary to *peal back the lid* and *look inside*. It is only then that it becomes apparent how the individual components of a DSR system can be viewed not as separate entities, but as the various organs of a complete body, and how optimal performance of such a system can be obtained.

Introduction

A complete system for distant speech recognition (DSR) typically consists of several distinct components. Among these are:

1. An array of microphone for far-field sound capture;
2. An algorithm for tracking the positions of the active speaker or speakers;
3. A beamforming algorithm for focusing on the desired speaker and suppressing noise, reverberation, and competing speech from other speakers;
4. A recognition engine to extract the most likely hypothesis from the output of the beamformer;
5. A speaker adaptation component for adapting to the characteristics of a given speaker as well as to channel effects;
6. Postfiltering to further enhance the beamformed output.

Moreover, several of these components are comprised of one or more subcomponents: The speaker tracking system will typically have a component for measuring the time delays of arrival between sensor pairs based, for example, on the generalized cross-correlation, adaptive eigenvalue decomposition, or mutual information. The tracking system will usually also contain a subcomponent, such as some form of Bayesian filter, for synthesizing these instantaneous measurements into a continuous time series of speaker position estimates. Similarly, beamforming will typically be performed in the frequency or subband domain. Hence, the beamforming component requires an analysis filter bank to convert the time signals from each sensor into subband samples. Thereafter these samples will be optimally combined using some optimization criterion, which is usually subject to a distortionless constraint in the look direction, and the final set of subband samples will be transformed back into the time domain using a synthesis filter bank. Finally, the recognition engine will typically make multiple passes over the output of the beamformer, performing speaker and channel adaptation between each pass. In distant ASR scenarios where the word error rates of the initial passes are especially high, word lattices generated during a prior pass are often used to estimate speaker and channel adaptation parameters for the next pass.

While it is tempting to develop each component in isolation, experience has proven that such an approach cannot lead to optimal performance. This follows from several causes: Firstly, the effect of each component can only be judged in terms of final system performance, which is most often judged in terms of word error rate (WER), but may also be measured in terms of dialogue completion rate, or user satisfaction. It often happens that the relatively simple metrics, such as signal-to-noise ratio (SNR), used to measure the performance of signal processing algorithms do not correlate well with metrics such as WER. Secondly, the several components of the complete system interact in ways that are often neither simple nor direct. An algorithm that is "locally" optimal may have disastrous consequences for another component "downstream". Finally, each component of the system typically has one primary output. Each component may well, however, preserve internal state that is potentially useful as side information for one or more other components of the system.

In this work, we will discuss several examples of the interactions between the individual components of a DSR system. In addition, we will describe the synergies that become possible as soon as each component is no longer treated as a "black box". To wit, instead of treating each component as having solely an input and an output, it is necessary to *peal back the lid* and *look inside*. It is only then that it becomes apparent how the individual components of a DSR system can be viewed not as separate entities, but as the various organs of a complete body, and how optimal performance of such a system can be obtained.

Box 1: Human Speech

In our understanding, the first component of a DSR system is the signal of interest, namely, the speech of the desired speaker. We refer to human speech as a black box *not* because nothing is known about it. All to the contrary: A *great deal* is known about human speech, but this in-

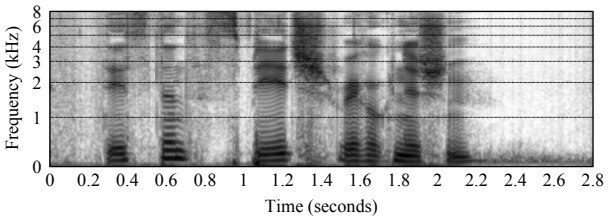

Figure 1: Spectrogram of typical utterance. From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

formation is very seldom used when developing the array processing components of a DSR system.

What then is known about human speech? First of all, speech is highly non-stationary. This is clearly evident from the the time-frequency plot or *spectrogram* shown in Figure 1. Secondly, speech is very sparse in the subband domain, this is also clearly evident from the spectrogram in Figure 1. Despite the fact that two speakers speak simultaneously, it is highly unlikely that their utterances will simultaneously have significant spectral content in the same subband. This is why binary masking works comparably well for speech separation as the fanciest beamforming algorithm. Thirdly, speech is largely, but not exclusively, periodic. This fact is accounted for by the well-known source-filter model of speech [37, §2.2.1], wherein there are two sources of excitation: a pulse train to model voiced segments of speech such as vowels, and a noise generator to model unvoiced segments such as fricatives. The sparseness of speech in the subband domain is largely due to the *overtone series* associated with periodic or voiced speech. The fundamental frequency, typically denoted as f_0, is the rate at which the vocal cords open and close. Because of this periodicity, there will be a great deal of energy in those subbands that fall directly on an integer multiple or *harmonic* of f_0. Between the harmonics, however, there will be very little energy in the spectral domain.

The entire field of *independent component analysis* (ICA) is founded on the assumption that all information-bearing signals are *not* Gaussian-distributed [12]. Briefly, the reasoning for this is grounded on two points:

1. The *central limit theorem* states that the pdf of the sum of independent random variables (r.v.s) will approach Gaussian in the limit as more and more components are added, *regardless* of the pdfs of the individual components. This implies that the sum of several r.v.s will be closer to Gaussian than any of the individual components. Thus, if the original independent components comprising the sum are sought, one must look for components with pdfs that are the *least* Gaussian.

2. The *entropy* for a continuous complex-valued r.v. Y, is defined as

$$H(Y) \triangleq -\mathscr{E}\{\log p_Y(v)\} = -\int p_Y(v)\log p_Y(v)dv, \quad (1)$$

where $p_Y(.)$ is the pdf of Y. Entropy is the basic measure of information in *information theory* [10]. It is well known that a Gaussian r.v. has the highest entropy of all r.v.s with a given variance [10, Thm. 7.4.1], which also holds for complex Gaussian r.v.s [26, Thm. 2]. Hence, a Gaussian r.v. is, in some sense, the least *predictable* of all r.v.s. Information-bearing signals contain structure that makes them more predictable than Gaussian r.v.s. Hence, if an interesting signal is sought, one must once more look for a signal that is *not* Gaussian.

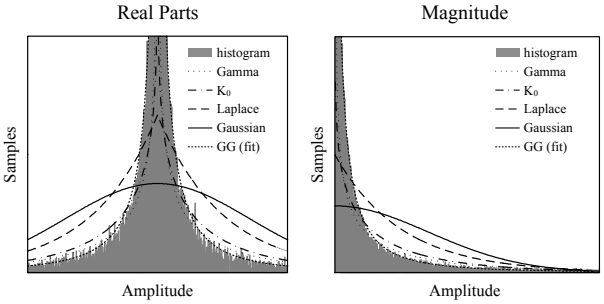

Figure 2: Histogram of real parts or magnitude of subband components and the likelihood of pdfs. From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

The fact that the pdf of speech is super-Gaussian has often been reported in the literature [14, 22, 8]. Noise, on the other hand, is more nearly Gaussian-distributed. In fact, the pdf of the sum of several super-Gaussian r.v.s. becomes closer to Gaussian. Thus, a mixture consisting of a desired signal and several interfering signals can be expected to be nearly Gaussian-distributed.

The super-Gaussianity of human speech is clearly evident from the plots of the probability density functions (pdfs) and histograms displayed in Figure 2. As is well known, super-Gaussian signals display infrequent, but large deviations from their mean values, and thus exhibit "spikey" and "heavy-tailed" characteristics. In comparison with the normal distribution, super-Gaussian pdfs have more probability mass near their means, less probability mass at intermediate values of their arguments, and far more probability mass in their tails; i.e., far away from their means. As alluded to above, the super-Gaussianity of human speech is connected with its sparsity in the subband domain. Each subband sample is very, very often nearly zero, but then briefly far-removed from zero.

Box 2: Realistic Acoustic Environments

The second component of a DSR system is the channel, namely, the real acoustic environment through which the desired signal propagates. We refer to the real acoustic environment as a black box for much the same reason as we have done so with speech, to wit, although much is known about them, this information is all too seldom used to build effective DSR systems.

Hence, it is also worthwhile to summarize what is known about the acoustic environments through which speech propagates. To state the obvious, these environments are characterized by the twin distortions of noise and reverberation [37, §2.4], which, as is apparent from Figures 3 and 4 respectively, have very different effects on speech. As shown in Figure 3, noise tends to "fill in" the low-energy regions of speech in the time-frequency domain, and thus has a *masking* effect. Reverberation, as shown in Figure 4, causes a temporal "smearing" of speech in the time-frequency domain; i.e., spectral content from one time instant is smeared into the following time instants.

An *echo* is a single reflection of a sound source, arriving some time after the direct sound. It can be described

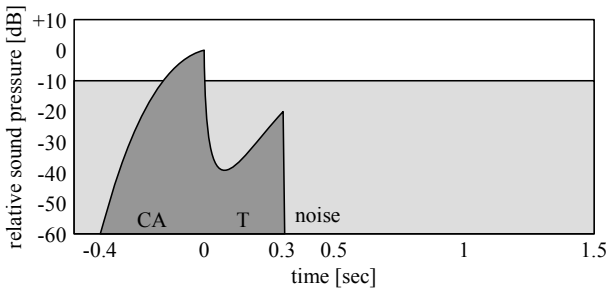

Figure 3: Simplified plot of relative sound pressure vs. time for an utterance of the word *cat* in additive noise. From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

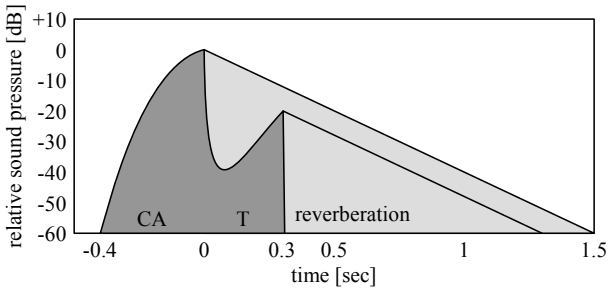

Figure 4: Simplified plot of relative sound pressure vs. time for an utterance of the word *cat* in a reverberant environment. From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

as a wave that has been reflected by a discontinuity in the propagation medium, and returns with sufficient magnitude and delay to be perceived as distinct from the sound arriving on the direct path. The human ear cannot distinguish an echo from the original sound if the delay is less than 1/10 of a second. This fact implies that a sound source must be more than 16.2 meters away from a reflecting wall in order for a human to be perceived an audible echo. *Reverberation* occurs when, due to numerous reflections, a great many echoes arrive nearly simultaneously so that they are indistinguishable from one another. Large chambers—such as cathedrals, gymnasiums, indoor swimming pools, and large caves—are good examples of spaces having reverberation times of a second or more and wherein the reverberation is cleanly audible. Those sound waves reaching the ear or microphone by the various paths which can be separated into three categories:

- *direct wave*
  The direct wave is the wave which reaches the microphone on a direct path. The time delay between the source and its arrival on the direct path can be calculated from the sound velocity and the distance from source to microphone. The frequency dependent attenuation of the direct signal is negligible [3].

- *early reflections*
  Early reflections arrive at the microphone on an indirect path within approximately 50 to 100 ms after the direct wave and are relatively sparse. There are frequency dependent attenuations of these signals due reflections from surfaces.

- *late reflections*
  Late reflections are numerous reflections that follow one another so closely that they become indistinguishable from one another and result in a diffuse noise field. The degradation becomes frequency dependent as the air attenuation [3] becomes more significant due to the greater distance that the sound must travel and the frequency dependence of the reflecting surfaces.

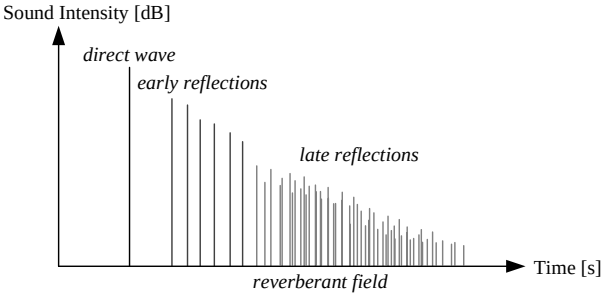

Figure 5: Early and late reflections of an impulse (direct wave). From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

A detailed pattern of the different reflections is presented in Figure 5. Note that this pattern changes drastically if either the source or the microphone moves, or the room impulse changes when, for example, a door or window is opened.

Clearly, the direct signal is most useful for ASR, but Nishiura *et al* [27] found that the early reflections up to 12.5 ms also improve ASR performance. This is significantly shorter, however, than the 50 ms time frame wherein early reflections were found to improve human recognition accuracy [18]. From a statistical point of view, the primary effect of noise and reverberation is to make the super-Gaussian subband samples of speech more nearly Gaussian. This can be clearly seen in the histograms displayed in Fig. 6. These figures demonstrate that both noise and reverberation cause probability mass to move out of the central spike and tail and into the intermediate regions of the pdf, which is to say, the pdf becomes more nearly Gaussian.

Finally, the realistic acoustic environments in which speech propagates typically contain *other speakers*. Hence, overlapping speech is a common phenomenon. This is why there has recently been great interest in the automatic recognition of overlapping speech [14, 23], which

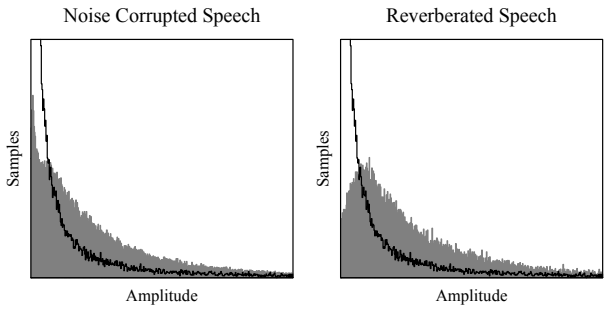

Figure 6: Histograms of subband magnitude of clean speech (black line) and noise corrupted or speech corrupted with reverberation. From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

is a source separation task *par excellence*. There has also been substantial effort devoted to the collection and annotation of corpora, such as the *Multi-Channel Wall Street Journal Audio Visual Corpus* (MC-WSJ-AV) [20], of overlapping speech. From a statistical point of view, the effect of overlapping speech is much like that of noise and reverberation: It tends to make a highly super-Gaussian signal more nearly Gaussian.

Box 3: Beamforming

Traditional beamforming algorithms are based on the minimization of some quadratic objective function—such as the variance of the beamformer output, mean square error, or signal-to-noise ratio—subject to a distortionless constraint. Recent research has revealed, however, that such criteria are *not* optimal for beamforming on human speech in realistic acoustic environments [15, 16, 29]. Once more, it is necessary to open the black box, and not simply use a generic algorithm while understanding neither the principles on which it is based nor its inherent limitations.

Given the observation that subband samples of speech become more nearly Gaussian when corrupted by noise or reverberation, it seems reasonable to consider restoring the orginal super-Gaussian statistical characteristics of speech as a means of combating these twin banes. *Negentropy* and *kurtosis* are two well-known measures of non-Gaussianity. The negentropy of a complex-valued r.v. Y is defined as

$$J(Y) \triangleq H(Y_{\text{gauss}}) - H(Y) \tag{2}$$

where Y_{gauss} is a Gaussian variable which has the same variance σ_Y^2 as Y. The entropy of Y_{gauss} can be expressed as

$$H(Y_{\text{gauss}}) = \log\left|\sigma_Y^2\right| + 2\left(1 + \log 2\pi\right). \tag{3}$$

From (1) and (2), it is clear that a super-Gaussian pdf is required to evaluate the negentropy of a random variable. The generalized Gaussian (GG) pdf is well-known and finds frequent application in the field of ICA. Moreover, it subsumes the Gaussian and Laplace pdfs as special cases. The GG pdf for a real-valued r.v. y with zero mean can be expressed as

$$p_{\text{GG}}(y) = \frac{1}{2\Gamma(1+1/p)A(p,\hat{\sigma})} \exp\left[-\left|\frac{y}{A(p,\hat{\sigma})}\right|^p\right], \tag{4}$$

where p is the *shape factor*, $\hat{\sigma}$ is the *scale factor* which controls how fast the tail of the pdf decays, and

$$A(p,\hat{\sigma}) \quad = \quad \hat{\sigma}\left[\frac{\Gamma(1/p)}{\Gamma(3/p)}\right]^{1/2}. \tag{5}$$

In (5), $\Gamma(.)$ is the gamma function. Note that the GG with $p = 1$ corresponds to the Laplace pdf, and that setting $p = 2$ yields the Gaussian pdf, whereas in the case of $p \to +\infty$ the GG pdf converges to a uniform distribution. Moreover, this pdf can readily be generalized for circular complex random variables [37, §B.6.1].

The *excess kurtosis* or simply *kurtosis* of a r.v. Y with zero mean, defined as

$$\text{kurt}(Y) \triangleq \mathscr{E}\{Y^4\} - 3(\mathscr{E}\{Y^2\})^2, \tag{6}$$

is, like negentropy, a measure of how *non-Gaussian* Y is [12]. The Gaussian pdf has zero kurtosis; pdfs with positive kurtosis are *super-Gaussian*; those with negative kurtosis are *sub-Gaussian*. From observed samples, we can approximate (6) as

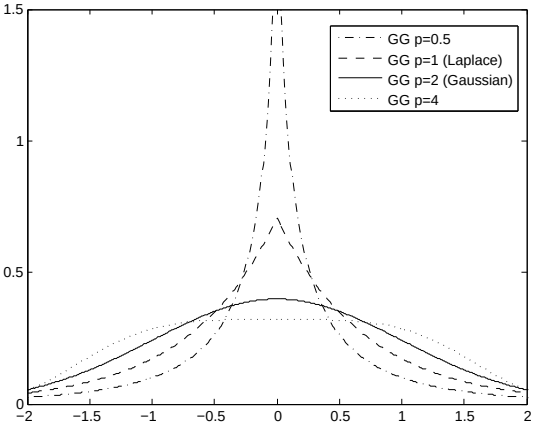

Figure 7: The generalized Gaussian (GG) pdfs.

$$\text{kurt}(Y) \approx \frac{1}{T}\sum_{t=0}^{T-1} Y_t^4 - 3\left(\frac{1}{T}\sum_{t=0}^{T-1} Y_t^2\right)^2. \tag{7}$$

Note that the empirical kurtosis measure requires no knowledge of the actual pdf of subband samples of speech, which is its primary advantage over negentropy as a measure of non-Gaussianity. The primary disadvantage of kurtosis with respect to negentropy is that it has been shown to yield worse beamforming performnance [16].

The kurtosis of the GG pdf can be controlled by adjusting the shape parameter p. Fig. 7 shows plots of the GG pdf with the same scaling factor $\hat{\sigma}^2 = 1$ and various shape factors, $p = 0.5, 1, 2, 4$. As p becomes smaller, the kurtosis of the GG increases. From the figure, it is clear that a smaller shape parameter yields a pdf with a spikier peak and heavier tail. As is clear from Fig. 7, as the kurtosis increases, the pdf becomes more spikey and heavy-tailed. It is apparent from Fig. 2 that the GG pdf is potentially far better-suited for modeling the statistics of subband samples of speech than other well-known pdfs. Moreover, the scale $\hat{\sigma}$ and shape p factors of the GG pdf can be efficiently estimated from training data [37, §13.5.2]. In all cases, the shape factors estimated from training data were $p < 2$, indicating positive kurtosis; i.e., super-Gaussian subband statistics.

In a recent overview paper, McDonough and Wölfel [25] posed three questions to the joint ASR and acoustic array processing communities:

- Why are the specific characteristics of speech so often ignored by the developers of ICA and beamforming algorithms? While the proposal of Seltzer [30] for incorporating a HMM directly into a beamforming algorithm is well-known, most ICA algorithms are formulated to work for any conceivable source [12], and most beamforming algorithms take no specific account of the characteristics of human speech.

- Why do developers of ICA algorithms consistently ignore geometric information?

- Why do developers of beamforming algorithms consistently ignore the use of higher order statistics?

The *maximum negentropy beamformer* (MNB) proposed by Kumatani *et al* [15] was intended to be a partial answer to these questions. These authors proposed to make use of the statistical effects of noise and reverberation on speech to perform improved beamforming on speech

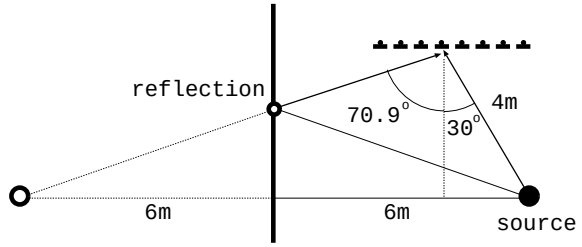

Figure 8: Configuration of a source, sensors, and reflective surface for simulation.

captured with far-field sensors. To wit, the measures of non-Gaussianity mentioned above were used to replace the minimum variance optimization criterion used in conventional beamforming. As is clear from the prior section, by adjusting the active weight vector of a beamformer in generalized sidelobe canceller (GSC) configuration so as to obtain an output signal that is maximally super-Gaussian, it is possible to suppress the detrimental effects of noise and reverberation [15].

The MVDR and MMSE beamformers attempt to minimize output power subject to a distortionless contraint [37, §13.3.3]. While such a constrained optimization criterion can suppress interference, it is also susceptible to the signal cancellation problem, whereby the desired signal is itself attenuated [35]. In contrast to such beamformers, the MNB attempts not only to eliminate interference signals but also *strengthen* those reflections from the desired source, as both behaviors tend to make the final signal more non-Gaussian. Of course, any reflected signal would be delayed with respect to the direct path signal. Such a delay would, however, manifest itself as a phase shift in the subband domain, and could thus be removed through a suitable choice of active weight vector. Hence, the MNB offers the possibility of steering both nulls *and* sidelobes; the former towards the undesired signal and its reflections, the latter towards reflections of the desired signal. Moreover, the MNB is not susceptible to the signal cancellation problem.

In order to verify that the MNB forms sidelobes directed towards the reflection of a desired signal, Kumatani *et al* [15] conducted experiments with a simulated acoustic environment. As shown in Fig. 8, those authors considered a simple configuration with a sound source, a reflective surface, and a linear array of eight microphones positioned with 10 cm inter-sensor spacing. Actual speech data were used as a source in this simulation, which was based on the *image method* [1]. White Gaussian noise was added to the data of each microphone. Kumatani *et al* assumed that the speed of sound is 343.74 meters per second and that the reflection coefficient is 0.7. Fig. 9 shows beam patterns at f_s = 150 Hz, f_s = 650 Hz and f_s = 1600 Hz obtained with a delay-and-sum (D&S) beamformer, the MVDR beamformer and the MNB with the GG pdf. The weights of the MVDR beamformer were optimized for isotropic (diffuse) noise in the simulation [4].

It is clear from Figure 9 that the MN beamformer emphasizes the reflection from the desired source, whereas the MVDR optimized for the diffuse noise field is unaffected by its presence. It is also apparent from Figure 9 (a) that the MVDR and MN beamformers can suppress interferences at low frequencies, where the suppression performance of the delay-and-sum beamformer is poor. Kurtosis was recently proposed as an alternate optimization criterion for beamforming by Kumatani *et al* [16].

Box 4: Recognition Engine

As illustrated in this section, the ASR engine can no more be treated as a black box than any of the other components in a complete DSR system.

In order to train a triphone acoustic model for DSR exerperiments, Kumatani *et al* [15] used 30 hours of American WSJ and the 12 hours of Cambridge Wall Street Journal (WSJ) data. Acoustic models estimated with two different HMM training schemes were used for several decoding passes: conventional maximum likelihood (ML) HMM training [7, §12], and speaker-adapted training under a ML criterion (ML-SAT) [2]. The baseline system was fully continuous with 1,743 codebooks and a total of 67,860 Gaussian components. Four decoding passes were performed on the waveforms obtained with various beamforming algorithms. Each pass of decoding used a different acoustic model or speaker adaptation scheme. Speaker adaptation parameters were estimated using the word lattices generated during the previous pass, as in [31]. A description of the four decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model.

2. Estimate vocal tract length normalization (VTLN) [34] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [9] for each speaker, then redecode with the conventional ML acoustic model.

3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [19] parameters for each speaker, then redecode with the conventional ML acoustic model.

4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SAT model.

A description of the fast on-the-fly technique used to dynamically compose two weighted finite-state transducers during the recognition passes, thereby enabling the application of a full trigram language model, is given in [24].

Kumatani *et al* [15] successfully applied a negentropy criterion to the acoustic beamforming problem for DSR. The advantages of incorporating higher-order statistics into beamforming algorithms through the use of non-Gaussian pdfs are illustrated by the results shown in Table 1, all of which were obtained on the single speaker portion of the MC-WSJ-AV corpus [20]. To wit, the MNB using the GG pdf proved more effective than the adaptive beamformers based both on the minimum mean square error (MMSE) criterion [37, §13.3.5] as well as the recently-proposed generalized eigenvector (GEV) blocking matrix [33] criterion. What is also clear from Table 1 is that the "conventional" ASR techniques, in particular speaker and channel adaptation based on word lattices, are also extraordinarily effective in reducing WER on DSR tasks. In particular, on the best reported case involving the MNB with the GG pdf, speaker and channel adaptation reduced WER from 75.1% on the first unadapted pass to 13.2% on the final pass with the strongest adaptation and acoustic model. Such is the reason that all state-of-the-art DSR systems use such techniques. For reference, the WER results obtained with the delay-and-sum (DS) beamformer, as well as with a single

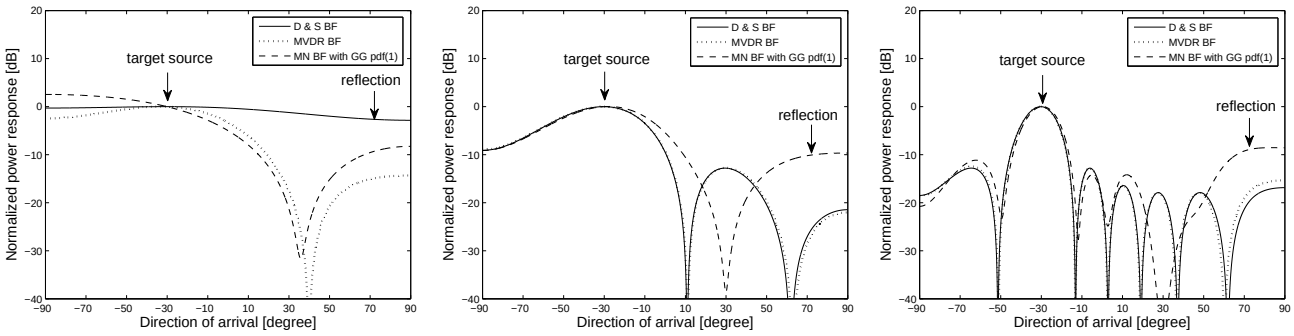

Figure 9: Beam patterns produced by a delay-and-sum beamformer, the MVDR beamformer and the MN beamforming algorithm using a spherical wave assumption for (a) f_s = 150 Hz, (b) f_s = 650 Hz and (c) f_s = 1600 Hz.

Table 1: Word error rates for each beamforming algorithm after every decoding pass.

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
D&S BF	79.0	38.1	20.2	16.5
MMSE BF	78.6	35.4	18.8	14.8
GEV BF	78.7	35.5	18.6	14.5
MN BF with GG pdf	75.1	32.7	16.5	13.2
SDM	87.0	57.1	32.8	28.0
CTM	52.9	21.5	9.8	6.7

distant microphone (SDM) and close-talking microphone (CTM) are also reported. The latter result indicates that further advances in beamforming are required to match the performance achieveable with a CTM using far-field sensors. It is worth noting that the best result of 13.2% in Table 1 is significantly less than half the word error rate reported elsewhere in the literature on this DSR task [20].

Buchner *et al* [5] have remarked that their TRINICON algorithm was the first blind source separation algorithm to simultaneously take into account 1) non-whiteness, 2) non-stationarity, and 3) non-Gaussianity. Kumatani *et al* [15] ignored the non-stationary nature of speech, but proved effective nonetheless, doubtless because this approach takes into the account the geometric information available from knowledge of the configuration of the sensor array as well as the location of the speaker.

The LIMA-BEAM algorithm proposed by Seltzer *et al* [30] is well-known. This algorithm is based on performing *maximum likelihood* (ML) beamforming in the log-spectral domain. The results reported by Raub *et al* [28] suggested that ML HMM beamforming was more effective when applied in the cepstral rather than in the log-spectral domain.

Is either the log-spectral or cepstral domain really the best domain for beamforming? For the purposes of beamforming, both have inherent drawbacks: To wit, in both cases, the fine frequency resolution required for beamforming has been smeared out through the application of one or more of linear prediction, Mel filter bank, and cepstral truncation. Similarly, in both domains, all phase information has been discarded. It should not be forgotten that beamforming in the subband domain is, to a first order, simply a linear phase shift. ASR and beamforming have different requirements with respect to time and frequency resolution: ASR requires good time resolution, but frequency resolution is unimportant. With beamforming, the opposite is true.

As seen in Figure 1, speech is a highly non-stationary signal. Hence, we can perhaps reinterpret Seltzer's intuition in formulating the LIMA-BEAM algorithm as follows: How would it be possible to model the non-stationarity of speech with a HMM for the purpose of adaptive beamforming? Given the good initial results of Kumatani *et al* [15], the demonstrated super-Gaussian nature of speech, and the desirability of performing beamforming in a domain where phase information is preserved, we might be lead to follow Rauch *et al* [29] in proposing the following solution: Use Kumatani's notion of beamforming under a maximum negentropy criterion, but, during a second pass, perform HMM alignment and extract the short-term power spectral density (PSD) of the speech from the adapted cepstral means of an *auxiliary model*. The auxiliary model would have the same state clustering as the main HMM used for recognition, but only a single Gaussian per codebook, and that with only static cepstral features. The auxiliary model is trained not with a normal cepstral front-end wherein a Hamming window is followed by a DFT, Mel filter bank and discrete cosine transform, but instead with the same uniform DFT analysis filter bank used for beamforming [17]. The value of the PSD at a given frequency ω_c is equivalent to the instantaneous variance σ_k^2 of the output of the kth subband with center frequency $\omega_k = \omega_c$. Hence, the instantaneous scale factor $\hat{\sigma}_k$ of the GG pdf used for maximum negentropy beamforming would be given by $\hat{\sigma}_k = \sqrt{\sigma_k^2}$. After Viterbi alignment of the auxiliary model with the subband samples from the analysis filter bank, the value of σ_k^2 for a particular analysis frame can be recovered by inverting the ceptral feature extraction process on the time-aligned cepstral mean from the auxiliary model in order to obtain the corresponding PSD. The PSD for the kth subband is in fact σ_k^2. The development of such an algorithm was the subject of an IWAENC 2008 contribution [29].

Figure 10 shows an example of the reconstructed log PSD for one frame of a test utterance beamformed with the MNB. The average PSD value is compared to the original PSD and the PSD reconstructed from the 13 original cepstral coefficients, as well as that reconstructed from the 13-coefficient cepstral mean in the HMM state aligned with this frame. We observe that the HMM-based recon-

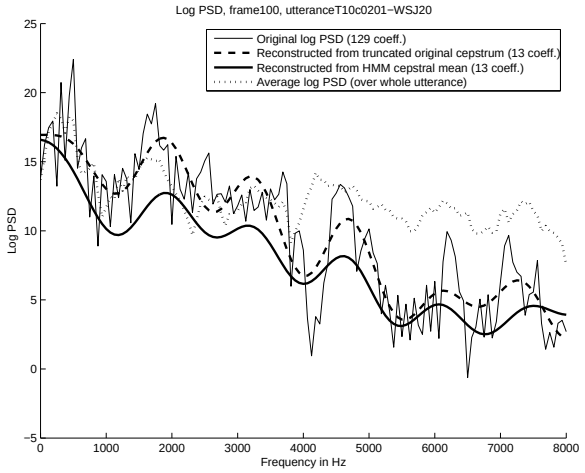

Figure 10: *Original and reconstructed log PSD of a test utterance.*

struction approximates the spectral envelope well in the log PSD domain. The PSD obtained by averaging over the entire utterance, on the other hand, models only the long-term *spectral tilt*, and thus does not capture the non-stationarity of human speech.

Table 2: WERs after every decoding pass.

Beamforming Algorithm	Pass (%WER)			
	1	2	3	4
D&S BF	80.1	39.9	21.5	17.8
MNB, global variance	75.3	34.8	18.2	**14.6**
HMM-MNB	74.9	32.7	16.9	**13.6**
HMM-MNB, oracle	75.0	33.7	17.2	14.1
SDM	87.0	57.1	32.8	28.0
CTM	52.9	21.5	9.8	6.7

Table 2 shows the WERs for every beamforming algorithm reported by Rauch *et al* [29]. The table entry marked "MNB, global variance" corresponds to the algorithm proposed by Kumatani *et al* [29]. As references, WERs in recognition experiments on speech data recorded with the single distant microphone (SDM) and with the close-talking microphone (CTM) are also given in Table 2. We observe that the HMM-MNB performs better than the global variance baseline on all passes, with a 1% absolute gain on the final pass. As the initial decoding passes will not always give correct results, we also provide results for an oracle experiment with optimistic HMM alignments; that is, alignments obtained with the correct transcriptions. The oracle results also proved better than the baseline, but worse than the non-oracle case. This may seem surprising at first since the real transcriptions could be expected to lead to more accurate speech modeling. We suspect that the superior performance of the non-oracle case is due to the fact that the incorrect hypothesis has by definition a higher likelihood than the correct transcript, and hence represents a better match between the HMM acoustic models and the data. This leads, in turn, to a more accurate reconstruction of the PSD.

Table 3: Word error rates without post-filtering for every filter bank design algorithm after every decoding passes.

Filter bank	Pass (%WER)			
	1	2	3	4
FFT	88.5	71.1	58.8	55.5
PR	87.7	65.2	54.0	50.7
de Haan	88.7	68.2	56.1	53.5
Nyquist(M)	88.5	67.0	55.6	52.5
CTM	37.1	24.8	23.0	21.6

Box 5: Filter Banks

While subband adaptive filtering and beamforming undoubtedly require an analysis and (possibly) a synthesis filter bank (FB), not all filter banks are equally suited for such applications. In particular, filter bank designs based on maximal decimation using an aliasing cancellation strategy, although well-suited for subband coding applications, often perform poorly in subband adptive filtering and beamforming applications. Hence, the FB represents yet another black box whose contents must be carefully examined before optimal performance can be obtained. We will discuss examples where the choice of the wrong filter bank effectively masked gains produced by downstream components of the system, such as the beamforming and postfiltering algorithms.

Shown in Table 3 are the results of ASR experiments run on data from the AMI Speech Separation Challenge, Part II [20]. In this task, two speakers simultaneously read sentences from the 5,000 word vocabulary WSJ task, and the data is captured with two circular, eight-channel microphone arrays. The task is then to separate and automatically recognize the speech of the two speakers by any available means. The signal processing and ASR algorithms used to generate the results in Table 3 are described in [14, 23]. The decoding passes used were those described under Box 4: Recognition Engine.

For the initial set of experiments, no post-filter was used. As indicated in Table 3, the perfect reconstruction (PR) filter bank described in [32, §8] proved to provide the lowest WERs, better in fact than the plain vanilla FFT, as well as the de Haan and Nyquist(M) filter banks described in [6] and [17] respectively. This was a surprising result, in that the PR design is based on the concept of aliasing cancellation, whereby the aliasing that is perforce present in one subband is cancelled by the aliasing present in all others. The latter functions correctly only when arbitrary scale factors and phase shifts are *not* applied to the outputs of the individual subbands, which is precisely what happens during beamforming. Note that the FFT provided the very worst results of all, which is not surprising given that the simple FFT provides very poor stopband suppression [37, §3.4.1]. For reference, a comparable set of results obtained with data captured with a close-talking microphone (CTM) is also tabulated.

Table 4 shows results on the same task when a Zelinski postfilter [21] was used after beamforming. In this case, the PR filter bank provides significantly worse performance than both the de Haan and Nyquist(M) designs. In particular, it is clear that systems with de Haan and Nyquist(M) filter banks can reduce the absolute WER by about 5.0% compared to those with the PR filter banks.

Table 4: Word error rates with post-filtering for every filter bank design algorithm after every decoding passes.

Filter bank	Parameters		Pass (%WER)			
	M	D	1	2	3	4
PR	64	-	83.7	61.5	47.5	44.7
	512	-	84.6	60.5	47.6	44.4
de Haan	64	32	82.4	59.2	46.2	43.3
	256	128	82.0	60.5	44.7	42.0
	512	256	83.9	59.1	43.2	41.3
	512	128	81.6	58.9	43.2	40.3
	512	64	82.7	57.7	42.7	39.6
Nyquist(M)	64	32	80.7	57.0	44.3	42.0
	256	128	81.0	56.2	41.8	39.8
	512	256	84.1	58.6	43.4	40.6
	512	128	81.8	54.9	42.2	39.6
	512	64	81.4	56.5	42.6	40.3

This suggests that the PR filter bank is less suitable for adaptive processing, as we have suggested. This was not apparent, however, until a postfilter was added to the output of the beamformer. The combination of changing filter banks and adding the postfilter greatly improved overall system performance.

Box 6: Speaker Tracking System

As we will describe in this section, the internal state of a speaker tracking system can be very useful for determining when the search for the most likely word hypothesis in an ASR engine should begin and end. It also provides information for associating utterances with active speakers, which in turn is useful for speaker adaptation. To profit from this internal state, it is once more necessary to look beneath the lid of this particular black box.

A highly successful approach to speaker tracking [13] is based on feeding an observation $\mathbf{y}_k$ corresponding to a time delay of arrival (TDOA) directly into a Kalman filter (KF), where k is the current frame index. The state $\mathbf{x}$ of the KF is equivalent to the position of the active speaker in room coordinates. Let $\mathbf{x}_{k|k-1}$ denote the estimate of the state using all observations up to time $k-1$. The predicted observation is by definition

$$\hat{\mathbf{y}}_k \triangleq \mathbf{H}_k \mathbf{x}_{k|k-1},$$

where $\mathbf{H}_k$ is the observation matrix. Moreover, the innovation is defined as

$$\mathbf{s}_k \triangleq \mathbf{y}_k - \hat{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{H}_k \mathbf{x}_{k|k-1}$$

During its operation, a Kalman filter continously updates both predicted and filtered state estimation error covariance matrices [37, §4.3.1], denoted as $\mathbf{K}_{k|k-1}$ and $\mathbf{K}_k$ respectively. These matrices indicate the region of uncertainty wherein the speaker is likely to be found. Based on $\mathbf{K}_{k|k-1}$ and $\mathbf{H}_k$, the covariance matrix $\mathbf{S}_k$ of the innovation $\mathbf{s}_k$ must be calculated in order to calculate the Kalman gain $\mathbf{G}_k$. The update of the state estimate then proceeds according to

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{G}_k \mathbf{s}_k.$$

where $\mathbf{x}_{k|k}$ is the state estimate using all observations up to time k. Further details can be found in [37, §4.3.1].

The filtered state error covariance matrix $\mathbf{K}_{k|k-1}$ is typically small when the speaker has been speaking constantly for several seconds. As soon as a given speaker stops speaking, however, $\mathbf{K}_{k|k-1}$ grows rapidly and therewith region of uncertainty wherein the speaker is located. Hence, setting a threshold on the volume of $\mathbf{K}_{k|k-1}$ provides a natural *segmentation* that is potentially useful for ASR; i.e., it indicates both the beginning and end points of an utterance.

The *joint probabilistic data association filter* (JPDAF) is an extension of the Kalman filter that can simultaneously maintain several active tracks [37, §4.3.6]. Such a system has been used as the basis for an acoustic speaker tracking system capable of tracking several simultaneously active speakers [11]. A schematic of the operation of the JPDAF is shown in Fig. 11. In the figure, two active tracks are shown with current position estimates $\hat{\mathbf{x}}^{(i)}_{k|k-1}$ and $\hat{\mathbf{x}}^{(j)}_{k|k-1}$. As shown in the figure, all of the quantities defined above must be maintained for each active track. Moreover, the JPDAF is capable of using multiple observations $\{\mathbf{y}^{(m)}_k\}$ at each time step k, and making a probabilistic association between observations and active tracks.

As with the KF, the covariance matrices $\mathbf{K}^{(i)}_{k|k-1}$ and $\mathbf{K}^{(j)}_{k|k-1}$ shrink when the speaker associated with the corresponding track is active, and grow again as soon as that speaker ceases speaking. Hence, as with the simpler KF, the volume enclosed by each filtered state error covariance matrix provides a natural segmentation of the speech; the probability that speech is present decreases as the volume enclosed by the filtered state estimation error covariance matrix grows. Moreover, due the fact that each filtered state estimation error correlation matrix is associated with a speaker position, such a segmentation system is capable of indicating not only if speech is present, but also which speaker at which position in the room is currently active. Inasmuch as speakers typically move but slowly, this position information is a powerful cue as to the speaker's identity, which is in turn very useful for clustering the utterances of a single speaker for use in speaker adaptation.

Box 7: Post Filtering

Compensating for the effects of noise and reverberation are traditionally seen as separate problems. This is apparent from Table 5, which summarizes various speech enhancement techniques that have been proposed in the literature. In realistic acoustic environments, however, these two distortions are both invariably present. To achieve the full potential of speech enhancement, techniques must be developed which have the ability to suppress both kinds of distortions in a single framework, and thus represent yet another black box to be opened. A first step in this direction, as proposed by Wölfel [36], is through modeling additive and reverberant distortions in a single framework based on particle filters. The basic idea is to jointly track both distortions in a high dimensional space. The lower B dimensions represent the energy of additive distortions for each frequency bin ω_k, for $k = 1, 2, \ldots, B$, while higher dimensions represent reverberation energy. Instead of tracking the reverberation energy directly only the scale term of the reverberation estimate, which has a significantly lower variance, is tracked. The filter has now the freedom to decide how much "weight" is given to the different distortions. Table 6 presents the reductions in WER achieved by compensating for additive and reverberant distortions,

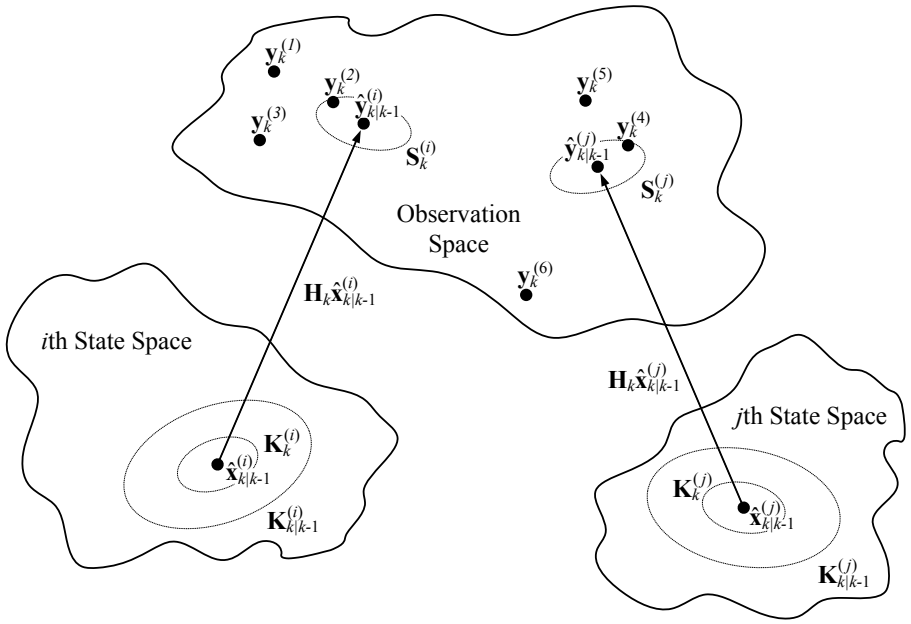

Figure 11: Schematic illustrating the operation of the joint probabilistic data association filter. From Wölfel and McDonough [37] (c) John Wiley & Sons, Ltd.

Method	Additive	Reverberation
blind deconvolution	no	yes
multi-step linear prediction	no	yes
harmonicity-based dereverberation	no	yes
work by Sehr *et al.*	no	yes
spectral subtraction	yes	no
Wiener filtering	yes	no
subspace algorithms	yes	no
vector Taylor series	yes	no
parallel model combination	yes	no
Bayesian filters, e.g. particle filters	yes	no
joint particle filter approach	joint	

Table 5: Standard techniques for speech enhancement.

either in isolation or jointly. It is apparent that enhancements from both techniques contribute to an increased performance of the overall system. A joint approach to speech enhancement, however, achieves the best overall performance. Thus, for optimal performance, enhancement techniques cannot compensate for one of the two distortions frequently encountered in distant speech recognition and leaving the compensation of the other kind of distortions to a black box.

To strengthen our arguments that the components in a DSR system, cannot be seen as black boxes and thus treated and optimized independently, we have performed experiments with different system configurations on a German large vocabulary speech recognition task. The task was similar to the experiments shown in Table 6, but in an environment with more stationary distortions. Comparing the results in Table 7 we observe that feature enhancement, here we used the joint approach as before, degraded if the filler model, everything else has been left identical, of the speech recognition system is not correctly adjusted. In the case, however, where the recognition setup is correctly adjusted the identical enhancement framework with identical setup does show significant improvements over the baseline of 5.4% absolute on the first pass and 0.4% absolute on the second pass. The small improvement over the baseline on the second pass can be explained by the ability of unsupervised acoustic model adaptation to compensate for stationary distortions.

Distance		150-200 cm		300-400 cm	
SNR		17 dB		10 dB	
Pass		1	2	1	2
Compensation		Word Error Rate %			
Additive	Reverberation				
no	no	18.6	14.0	45.4	28.6
yes	no	17.8	13.2	42.8	25.4
no	yes	17.7	13.4	39.2	23.9
yes	yes	17.7	13.3	38.3	23.3
joint		16.9	12.6	38.4	22.2

Table 6: Speech recognition experiments on single channel recordings with different speaker to microphone distances. From Wölfel [36].

Microphone		CTM		Distant	
SNR		24 dB		12 dB	
Pass		1	2	1	2
Optimized for	Compensation	Word Error Rate %			
CTM	no	12.7	12.4	48.8	28.0
CTM	yes	–	–	54.9	32.6
Distant	no	12.7	12.6	41.1	24.8
Distant	yes	–	–	35.7	24.4

Table 7: Speech recognition experiments on single channel recordings with different filler optimization.

Conclusions

In this work, we have investigated the specific properties and mutual interaction of all components of a complete DSR system. We began by examining the statistical characteristics of human speech, and then considered how these characteristics are altered when the speech propagates through an enclosed space. We then discussed how optimization criteria adopted from the field of independent component analysis, most noteably negentropy and kurtosis, can be profitably applied to acoustic beamforming for sources comprised of human speakers. Next we explained the importance of the ASR engine in a DSR system, and explored how information from such an engine together with an auxiliary HMM can be used to model the non-stationary of human speech, and thereby improve both beamforming and overall system performance. Thereafter, we investigated the decisive role of proper filter bank design in maximizing the performance of a DSR system. Finally, we described how a post filter based on a particle filter can be used to simultaneously compensate for the effects of noise and distortion, both of which are invariably present in any realistic acoustic environment. We also described the importance of proper modeling of noises in a DSR system.

The common thread running throughout this work has been the absolute necessity of considering the effectiveness of each component of a complete DSR not in isolation, but in the context of the whole. Moreover, we provided several examples where the internal state of one component can be effectively used as "side" information to enhance the performance of one or more other components, and thereby that of the entire system. Finally, we have presented results based either on acoustic simulations, or else on DSR experiments conducted on *real* acoustic data captured from *real* speakers and in *real* acoustic environments, thereby demonstrating the effectiveness of such a "wholistic" approach to the design of DSR systems. A more detailed and complete exposition of such a wholistic approach to DSR system design can be found in Wölfel and McDonough [37].

Acknowledgments

The authors gratefully acknowledge the financial support of the German Research Foundation (DFG) in connection with the International Research Training Network (IRTG) 715: "Language Technology and Cognitive Systems" and with Sonderforschungsbereich (SFB) 588: "Humanoid Robots—Learning and Cooperating Multimodal Robots". The authors are also grateful to John Wiley & Sons, Ltd. for permission to reproduce the images contained in this work.

References

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acou stics. *J. Acoust. Soc. Am.*, 65(4):943–950, April 1979.

[2] T. Anastasakos, J. McDonough, R. Schwarz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proc. ICSLP*, pages 1137–1140, 1996.

[3] H. E. Bass, H.-J. Bauer, and L. B. Evans. Atmospheric absorption of sound: analytical expression. *Jour. of ASA*, pages 821–825, 1972.

[4] M. Brandstein and D. Ward, editors. *Microphone Arrays*. Springer Verlag, Heidelberg, Germany, 2001.

[5] H. Buchner, R. Aichner, and W. Kellermann. Blind source seperation for convolutive mixtures: A unified treatment. In *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pages 255–289. Kluwer Academic, Boston, 2004.

[6] J. M. de Haan, N. Grbic, I. Claesson, and S. E. Nordholm. Filter bank design for subband adaptive microphone arrays. *IEEE Trans. Speech Audio Proc.*, 11(1):14–23, Jan. 2003.

[7] J. Deller, J. Hansen, and J. Proakis. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing, New York, 1993.

[8] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Transactions on Audio, Speech and Language Processing*, 15:1741–1752, 2007.

[9] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12, 1998.

[10] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, 1968.

[11] T. Gehrig, U. Klee, J. McDonough, S. Ikbal, M. Wölfel, and C. Fügen. Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters. In *Proc. Interspeech*, pages 2594–2597, 2006.

[12] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[13] Ulrich Klee, Tobias Gehrig, and John McDonough. Kalman filters for time delay of arrival–based source localization. *Journal of Advanced Signal Processing, Special Issue on Multi–Channel Speech Processing*, August 2005.

[14] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel. Adaptive beamforming with a minimum mutual information criterion. *IEEE Transactions on Audio, Speech and Language Processing*, 15:2527–2541, 2007.

[15] K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li. Adaptive beamforming with a

maximum negentropy criterion. In *Proc. Hands-Free Speech Communication and Microphone Arrays*, Trento, Italy, May 2008.

[16] K. Kumatani, J. McDonough, B. Rauch, P. N. Garner, W. Li, and J. Dines. Maximum kurtosis beamforming with the generalized sidelobe canceller. In *Proc. Interspeech*, September 2008.

[17] K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. Li. Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming. In *Proc. ICASSP*, 2008.

[18] H. Kuttruff. *Room Acoustics*. Elsevier Applied Science, 2000.

[19] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185, April 1995.

[20] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti. The multi-channel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments. In *Proc. ASRU*, pages 357–362, 2005.

[21] C. Marro, Y. Mahieux, and K. U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6:240–259, 1998.

[22] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Proc.*, 13(5):845–856, Sept. 2005.

[23] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow. To separate speech! a system for recognizing simultaneous speech. In *Proc. Machine Learning and Multi-modal Interfaces*, 2007.

[24] J. McDonough, E. Stoimenov, and D. Klakow. An algorithm for fast composition of weighted finite-state transducers. In *Proc. ASRU*, December 2007.

[25] J. McDonough and M. Wölfel. Distant speech recognition: Bridging the gaps. In *Proc. Hands-Free Speech Communication and Microphone Arrays*, 2008.

[26] F. D. Neeser and J. L. Massey. Proper complex random processes with applications to information theory. *IEEE Trans. Info. Theory*, 39(4):1293–1302, July 1993.

[27] T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama. Investigations into early and late reflections on distant–talking speech recognition toward suitable reverberation criteria. In *Proc. of Interspeech*, 2007.

[28] D. Raub, J. McDonough, and M. Wölfel. A cepstral domain maximum likelihood beamformer for speech recognition. In *Proc. Interspeech*, 2004.

[29] B. Rauch, K. Kumatani, J. McDonough, and D. Klakow. Hidden markov model beamforming with a maximum negentropy optimization criterion. In *Proc. International Workshop on Acoustic Echo and Noise Control*, September 2008.

[30] M. L. Seltzer, B. Raj, and R. M. Stern. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Trans. Speech Audio Proc.*, 12(5):489–498, September 2004.

[31] L. Uebel and P. Woodland. Improvements in linear transform based speaker adaptation. In *Proc. ICASSP*, 2001.

[32] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs, 1993.

[33] E. Warsitz, A. Krueger, and R. Haeb-Umbach. Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, U.S.A, 2008.

[34] L. Welling, H. Ney, and S. Kanthak. Speaker adaptive modeling by vocal tract normalizatio. *IEEE Trans. Speech Audio Proc.*, 10(6):415–426, 2002.

[35] B. Widrow, K. M. D., R. P. Gooch, and W. C. Newman. Signal cancellation phenomena in adaptive antennas: Causes and cures. *IEEE Transactions on Antennas and Propagation*, AP-30:469–478, 1982.

[36] M. Wölfel. A joint particle filter and multi-step linear prediction framework to provide enhanced speech features prior to automatic recognition. In *Proc. Hands-Free Speech Communication and Microphone Arrays*, Trento, Italy, May 2008.

[37] M. Wölfel and J. McDonough. *Distant Speech Recognition*. Wiley & Sons, Chichester, West Sussex, England, 2009.