frontiers
in Robotics and AI

MINI REVIEW
published: 07 July 2020
doi: 10.3389/frobt.2020.00085

A Survey on Probabilistic Models in Human Perception and Machines

Lux Li[1]*, *Robert Rehr*[2], *Patrick Bruns*[1], *Timo Gerkmann*[2] *and Brigitte Röder*[1]

[1] *Biological Psychology and Neuropsychology, University of Hamburg, Hamburg, Germany,* [2] *Signal Processing (SP), Department of Informatics, University of Hamburg, Hamburg, Germany*

Extracting information from noisy signals is of fundamental importance for both biological and artificial perceptual systems. To provide tractable solutions to this challenge, the fields of human perception and machine signal processing (SP) have developed powerful computational models, including Bayesian probabilistic models. However, little true integration between these fields exists in their applications of the probabilistic models for solving analogous problems, such as noise reduction, signal enhancement, and source separation. In this mini review, we briefly introduce and compare selective applications of probabilistic models in machine SP and human psychophysics. We focus on audio and audio-visual processing, using examples of speech enhancement, automatic speech recognition, audio-visual cue integration, source separation, and causal inference to illustrate the basic principles of the probabilistic approach. Our goal is to identify commonalities between probabilistic models addressing brain processes and those aiming at building intelligent machines. These commonalities could constitute the closest points for interdisciplinary convergence.

Keywords: signal processing, multisensory perception, audiovisual integration, optimal cue integration, causal inference, speech enhancement, automatic speech recognition, human psychophysics

OPEN ACCESS

Edited by:
Subhadeep Chakraborty,
The University of Tennessee, Knoxville, United States

Reviewed by:
Eric Wade,
The University of Tennessee, Knoxville, United States
Farshad Salimi Naneh Karan,
Independent Researcher, Denver, United States

****Correspondence:***
Lux Li
luxi.li@uni-hamburg.de

Specialty section:
This article was submitted to Sensor Fusion and Machine Perception, a section of the journal Frontiers in Robotics and AI

Received: *27 August 2019*
Accepted: *29 May 2020*
Published: *07 July 2020*

Citation:
Li L, Rehr R, Bruns P, Gerkmann T and Röder B (2020) A Survey on Probabilistic Models in Human Perception and Machines. Front. Robot. AI 7:85. doi: 10.3389/frobt.2020.00085

INTRODUCTION

Human perception and machine signal processing (SP) both face the fundamental challenge of handling uncertainty. Probabilistic models provide powerful tools for representing and resolving uncertainty (Rao et al., 2002). For example, a simple probabilistic model for estimating a speech signal from a noisy audio recording can be constructed as follows: The stimulus parameter of interest (e.g., the phoneme) is represented as a latent variable S. The existing information or expectation regarding S prior to the data observation is represented by the prior probability distribution, $p(S)$ ("*prior*"). The perceptual system's responses (often referred to as *measurements*) are usually stochastic: they fluctuate from trial to trial even when the stimulus remains constant. The conditional probability density function (PDF) of obtaining the measurements X given S is described by the likelihood function of S, $p(X|\ S)$ ("*likelihood*"). Probabilistic models commonly use the framework of *Bayesian inference*, which specifies how belief is optimally updated in light of new evidence. Computationally, this is achieved by applying the Bayes' theorem (Pouget et al., 2013; Ghahramani, 2015) to combine the likelihood and the prior to calculate the posterior probability distribution ("*posterior*"), $p(S\ |X)$:

$$p(S|X) = p(X|S)\, p(S)/p(X) \tag{1}$$

Signal reconstruction often requires a point-estimator for S. Three methods are commonly used. The maximum likelihood estimator (MLE) is the S value that maximizes the likelihood (Equation 2) or equivalently the log-likelihood, implying a uniform (flat) prior. The maximum *a-posteriori* (MAP) estimator can be seen as maximizing the likelihood after factoring in an informative prior (Equation 3) and is equal to the posterior mode. The minimum mean square error (MMSE) estimator is the *a-posteriori* expected value for S (Equation 4) and is equal to the posterior mean (Yuille and Bülthoff, 1996; Maloney, 2002).

$$\mathrm{MLE}: \hat{S} = arg \max_{s_i} p(X|S_i) \tag{2}$$

$$\mathrm{MAP}: \hat{S} = arg \max_{s_i} p(S_i|X) \tag{3}$$

$$\mathrm{MMSE}: \hat{S} = \int S_i \, p(S_i|X) dS_i \tag{4}$$

Similar probabilistic approaches are applied in sensory perception and machine SP for solving analogous problems, such as robust perception. However, although recent reviews have separately summarized probabilistic models in each of these disciplines (Kolossa and Häb-Umbach, 2011; Ma, 2012; Hendriks et al., 2013; Ursino et al., 2014), reviews that draw parallels between the models across the disciplines are lacking. Here, we will introduce and compare selective applications of probabilistic models in psychology, neuroscience, and machine SP, focusing on audio and audio-visual processing. We use the topics of speech enhancement, automatic speech recognition, audio-visual cue integration, and source separation as examples, because probabilistic models have played a particularly important role in advancing these research areas. We emphasize two important aspects of resolving uncertainty: noise reduction and source separation. While in recent years machine learning approaches have had a great impact in SP (Deng and Li, 2013; Padmanabhan and Premkumar, 2015), neuroscience (Yamins and DiCarlo, 2016), and cognitive science (Lake et al., 2017), here we highlight the commonalities between basic probabilistic models for machine and perceptual SP.

NOISE REDUCTION AND SPEECH ENHANCEMENT

Statistical approaches in speech enhancement for reducing background noise usually deal with single-channel signals, e.g., from a single microphone. The variance of a signal is generally understood as the power of the signal, and the PDFs characterize the coefficients of the digitized signals. Traditionally, the complex Fourier coefficients of the speech and noise components are modeled with a zero-mean Gaussian distribution [but later research suggests that super-Gaussian PDFs are more appropriate; see Lotter and Vary (2005), Martin (2005), and (Rehr and Gerkmann, 2018)], and the frequency bands are assumed to be statistically independent (Ephraim and Malah, 1984, 1985; Porter and Boll, 1984). The variances (i.e., the power) of the speech and noise coefficients are time-variant; therefore, the parameters must be continuously updated using adaptive power estimators. A common way to derive the estimators is by computing the MMSE between the true speech coefficients and the estimated coefficients, which leads to a linear filter known as the Wiener filter (Ephraim and Malah, 1984; Martin, 2001; Gerkmann and Hendriks, 2012). The Wiener filter has been adapted for multi-channel (e.g., multi-microphone array) processing (Krawczyk-Becker and Gerkmann, 2016), which additionally allows exploiting the spatial properties of sound (Kay, 1993; Balan and Rosca, 2002; Doclo et al., 2015). For multi-channel noise reduction, a well-known concept is the minimum-variance distortionless response (MVDR) beamformer. This beamformer minimizes the power of the output signal while ensuring that the sounds from the target speaker are not distorted or suppressed. The MVDR beamformer can be derived as the MLE of the speech coefficients if the background noise is assumed to follow a multivariate complex Gaussian distribution (Kay, 1993; Balan and Rosca, 2002).

Another classical probabilistic approach for estimating speech and noise coefficients is to use mixture models, most commonly Gaussian mixture models (GMMs) and hidden Markov models (HMMs) (Rabiner, 1989), with machine-learning methods (Ephraim, 1992; Burshtein and Gannot, 2002; Zhao and Kleijn, 2007; Chazan et al., 2016). The time-varying speech components are characterized by a sequence of discrete states related to the phonemes uttered by a speaker. Each state is described by a PDF linking it to the statistics of the observations. GMMs explicitly quantify the joint contributions of different states, whereas HMMs treat the states as latent variables that are related through Markov processes. The resulting estimator is a mixture of clean speech estimates from all possible combinations of available states; the states that best explain the observations have the strongest influence on the overall estimate. The advantage of a mixture estimator is that it takes into account all possible states and is more robust than basic MLEs.

Auditory systems of animals maintain robust neuronal representation of relevant sounds in noisy environments (Mesgarani et al., 2014). The dominant model for characterizing auditory neuronal responses is the spectrotemporal receptive field (STRF) (Zhao and Zhaoping, 2011; David, 2018; King et al., 2018). STRF is a linear filter that approximates the neuronal response at a given time as a linear weighted sum of the stimulus power at recent time points in different spectral channels (King et al., 2018). The weights can be viewed as a discrete-time version of the Wiener filter if they are estimated via the MMSE between the model output and the measured neuronal response, assuming Gaussian response noise with constant variance (Meyer et al., 2017). STRF is usually applied as part of a linear-nonlinear (LN) model—linear input followed by static nonlinear response generation (Chichilnisky, 2001; Paninski, 2003; Sharpee et al., 2004). However, standard STRF and LN models do not incorporate the highly nonlinear and dynamic neural processes which are important for noise robustness (for reviews, see Meyer et al., 2017; King et al., 2018). For example, auditory neurons adapt to stimulus statistics, such as the mean level and the contrast (i.e., the sound level variance) of recent sounds, and adjust their sensitivity accordingly; this adaptation enables efficient and robust neural coding (Fritz et al., 2003; David et al., 2012; Rabinowitz et al., 2013; Willmore et al., 2014, 2016; Lohse et al., 2020). STRF models extended with adaptive kernels (Rabinowitz et al., 2012) and other nonlinear features, such as input nonlinearity (Ahrens et al., 2008), synaptic depression (Mesgarani et al., 2014), gain normalization (Mesgarani et al., 2014), or top-down influence, such as feedback (Calabrese et al., 2011) and selective attention (Mesgarani and Chang, 2012), have been shown to better account for noise robustness. In addition, mixture-model approaches from SP (e.g., GMM) can be used to scale these models to higher-dimensional stimuli (Theis et al., 2013). In machine SP, machine-learning algorithms inspired by the nonlinear, adaptive, and/or top-down features of auditory neurons are being developed to improve speech enhancement (Ephraim, 1992; Hendriks et al., 2013; Lee and Theunissen, 2015; Rehr and Gerkmann, 2018, 2019). Future research could aim at building brain-inspired robust and flexible models to cope with various noise types, cluttered real-world data, and adversarial data.

AUDIO-VISUAL INTEGRATION MODELS IN A SINGLE-SOURCE SETUP

Probabilistic approaches have been extensively used for automatic speech recognition (ASR): the translation of audio signals into written text. Identifying the spoken words based only on the acoustic input signal can be challenging, especially if noise is present. Incorporating visual information (e.g., mouth shape, lip movement) can substantially improve ASR performance (Hennecke et al., 1996) in noisy environments, because visual features provide contextual and complementary (but additionally redundant) information about the audio scene and are insensitive to the acoustic background noise (Nefian et al., 2002). This approach is known as audio-visual speech recognition (AVSR). AVSR systems require dynamic models for optimal audio-visual (AV) integration. The performance of conventional HMMs, although being time-flexible, is limited by their strong restrictive assumptions, e.g., that the signal-generating system is a single process with few states and an extremely limited state memory (Brand et al., 1997). Nevertheless, a variety of HMM extensions have been proposed to better solve the AV fusion problem (Potamianos et al., 2003). One approach is to use a combination of feature fusion and decision fusion (Neti et al., 2000; Potamianos et al., 2003). *Feature fusion* applies fusion on the feature level; it trains a single HMM classifier on the concatenated vector of audio and visual features (Adjoudani and Benoît, 1996). *Decision fusion* applies fusion on the classifier output level; it linearly combines the likelihoods of audio-only and visual-only streams into a joint AV likelihood, using weights that capture the reliability of each sensory modality (Jain et al., 2000; Neti et al., 2000). Measures of reliability include the inverse variance (Hershey et al., 2004), signal-to-noise ratio (Adjoudani and Benoît, 1996; Hennecke et al., 1996), harmonics-to-noise ratio (Yumoto et al., 1982), or an equivalent index (Neti et al., 2000).

Two other extensions of HMMs are coupled HMMs (Brand et al., 1997; Abdelaziz et al., 2015) and factorial HMMs (Ghahramani and Jordan, 1997). These models have several advantages over conventional HMMs for AVSR: (1) they allow state asynchrony between the audio and visual components while preserving their natural correlation over time (Nefian et al., 2002; Abdelaziz et al., 2015), (2) they can model multiple interacting processes without violating the Markov condition (Brand et al., 1997), (3) the distributed state representations employed by these models allow automatic decomposition of superposed states (Ghahramani and Jordan, 1997), and (4) they are less sensitive to the initial conditions of parameters (Brand et al., 1997).

AVSR models are inspired by the human ability of using visual information to reduce auditory ambiguity (Schwartz et al., 2004). In human perception, a research topic related to AV fusion is generally known as *cue integration*. A *cue* is a sensory signal that bears information about the state of some stimulus property, e.g., identity or position. Psychophysical and neurophysiological studies have shown that the brain combines multiple cues both within and across sensory modalities to reduce uncertainty (for a review, see Fetsch et al., 2013). Computationally, to reduce uncertainty means to minimize the variance of perceptual estimates. One of the most well-known computational models for cue integration in psychophysics is the *forced fusion model* (**Figure 1A**), also known as the optimal cue integration model or the MLE model. This model proposes that a minimum-variance estimate for the target stimulus attribute S given multiple cues can be computed as the weighted linear sum of the MLEs for individual cues, and the weights are determined by each cue's relative reliability (Alais and Burr, 2004; Ernst and Bülthof, 2004; Rohde et al., 2015). A cue's *reliability* is defined as its inverse variance, $\frac{1}{\sigma_i^2}$, which is akin to how reliability is defined in a MVDR beamformer (Kay, 1993; Balan and Rosca, 2002). The forced fusion model assumes that the cues are *redundant*, i.e., they are regarding a single stimulus attribute and therefore should be completely integrated. Under the simplifying assumptions of a uniform prior $p(S)$ and independent Gaussian noises, the posterior $p(S \mid X_1, X_2, \ldots, X_\mathrm{n})$ is also a Gaussian, with its mean given by weighted summation:

$$\hat{S}_{opt} = \sum_{i=1}^{n} w_i \hat{S}_i \, , w_i = \frac{\frac{1}{\sigma_i^2}}{\sum_i^n \frac{1}{\sigma_i^2}} \tag{5}$$

where $\hat{S}_{opt}$ is the optimal combined estimate, $\hat{S}_i$ is the MLE for an individual cue i, and w_i is the weight determined by the relative reliability of cue i. These weights (w_i) minimize the variance of the combined estimate, and thus $\hat{S}_{opt}$ is a minimum-variance unbiased estimator for S [for a mathematical proof, see Colonius and Diederich (2018)]. This forced fusion model is analogous to the aforementioned fusion models used in multi-stream HMM for AVSR (Neti et al., 2000). The reliability-based weighting is similar to the stream weights that are determined by the inverse variance (Hershey et al., 2004). However, in the forced fusion model the weights are fixed, while in AVSR it has been shown that dynamic stream weights resulted in better performance (Meutzner et al., 2017). Furthermore, even in the seemingly simple case of fusing information from multiple microphones, the noise captured by individual microphones is typically correlated, especially in low frequencies. As a consequence, the

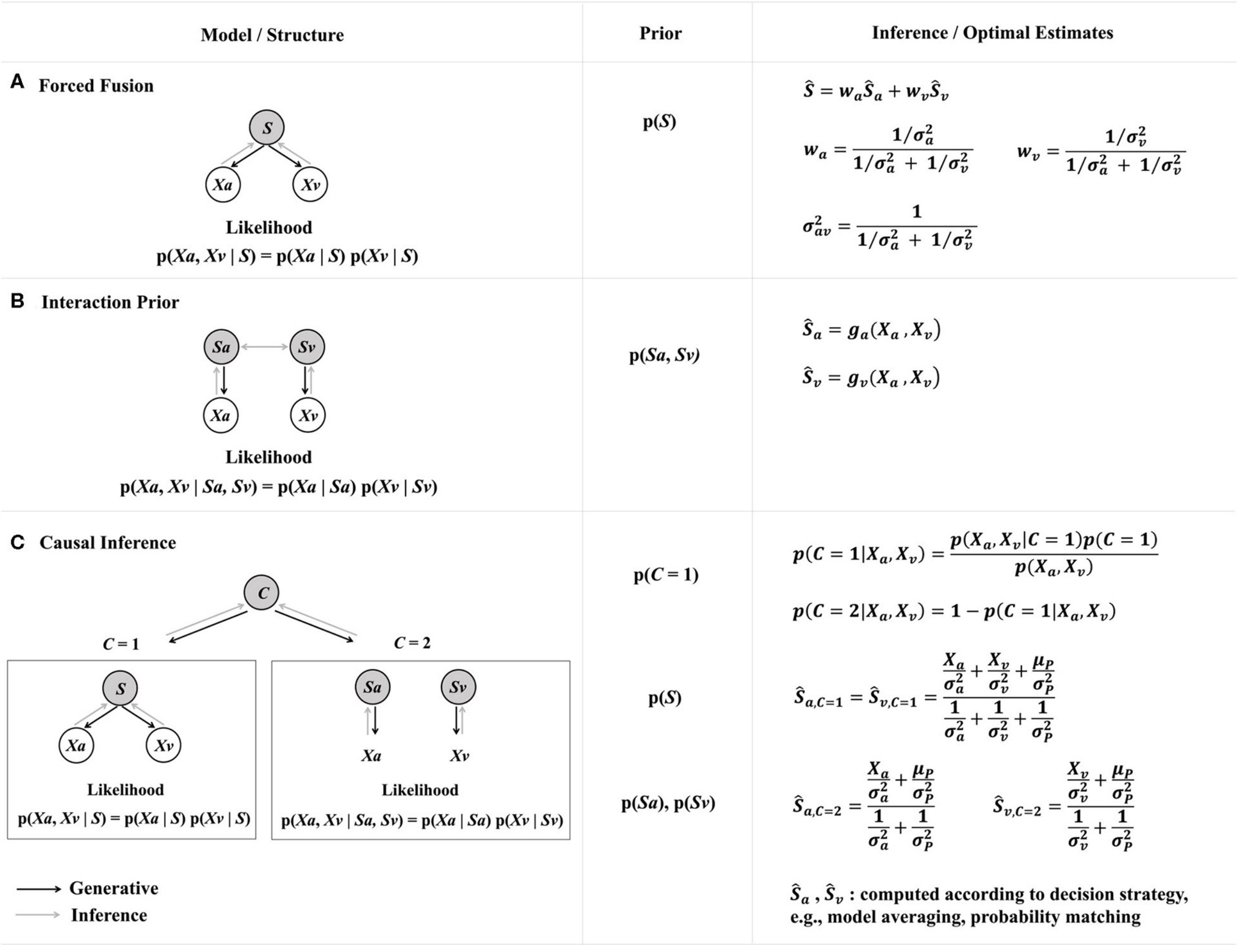
Model / Structure
Prior
Inference / Optimal Estimates
A Forced Fusion
S
Xa
Xv
Likelihood
p(Xa, Xv | S) = p(Xa | S) p(Xv | S)
p(S)
$\hat{S} = w_a\hat{S}_a + w_v\hat{S}_v$
$w_a = \frac{1/\sigma_a^2}{1/\sigma_a^2 + 1/\sigma_v^2}$
$w_v = \frac{1/\sigma_v^2}{1/\sigma_a^2 + 1/\sigma_v^2}$
$\sigma_{av}^2 = \frac{1}{1/\sigma_a^2 + 1/\sigma_v^2}$
B Interaction Prior
Sa
Sv
Xa
Xv
Likelihood
p(Xa, Xv | Sa, Sv) = p(Xa | Sa) p(Xv | Sv)
p(Sa, Sv)
$\hat{S}_a = g_a(X_a, X_v)$
$\hat{S}_v = g_v(X_a, X_v)$
C Causal Inference
C
C = 1
C = 2
S
Xa
Xv
Likelihood
p(Xa, Xv | S) = p(Xa | S) p(Xv | S)
Sa
Sv
Xa
Xv
Likelihood
p(Xa, Xv | Sa, Sv) = p(Xa | Sa) p(Xv | Sv)
p(C = 1)
p(S)
p(Sa), p(Sv)
$p(C=1|X_a,X_v) = \frac{p(X_a,X_v|C=1)p(C=1)}{p(X_a,X_v)}$
$p(C=2|X_a,X_v) = 1 - p(C=1|X_a,X_v)$
$\hat{S}_{a,C=1} = \hat{S}_{v,C=1} = \frac{\frac{X_a}{\sigma_a^2}+\frac{X_v}{\sigma_v^2}+\frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_a^2}+\frac{1}{\sigma_v^2}+\frac{1}{\sigma_P^2}}$
$\hat{S}_{a,C=2} = \frac{\frac{X_a}{\sigma_a^2}+\frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_a^2}+\frac{1}{\sigma_P^2}}$
$\hat{S}_{v,C=2} = \frac{\frac{X_v}{\sigma_v^2}+\frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_v^2}+\frac{1}{\sigma_P^2}}$
Generative
Inference
$\hat{S}_a$, $\hat{S}_v$: computed according to decision strategy, e.g., model averaging, probability matching

FIGURE 1 | Three probabilistic models for audio-visual cue integration in human psychophysics. Gray nodes depict the latent stimulus attribute S (e.g., identity or position) or the latent causal structure C. White notes depict the sensory measurements X in response to the sensory cues (a: auditory, v: visual). Left panel: The generative models and the underlying structures. The likelihood functions are derived under the assumptions that the auditory and visual cues are corrupted by independent Gaussian noise. Black arrows represent the direction of generative process, and gray arrows represent the direction of inference. Middle panel: A-priori knowledge. Right panel: Optimal estimates by Bayesian inference (adapted from Ursino et al., 2014 Box 1, copyright © 2014 Elsevier Ltd, and Shams and Beierholm, 2010 Figure 1, copyright © 2010 Elsevier Ltd; reused with permission). **(A)** Forced fusion model. The auditory and visual cues are assumed to have a common cause. The prior is usually assumed to be uniform, in which case this model is equivalent to an MLE. The optimal estimate is a linear weighted summation of unimodal MLEs, and the weights are the relative cue reliabilities (precision). This model describes complete cue integration (fusion). **(B)** Interaction prior model. The joint prior distribution $p(S_a, S_v)$ reflects the prior knowledge about the audio-visual correspondence in the environment. A common choice is a 2D Gaussian or Gaussian-mixture function with higher probabilities along the identity line $S_a = S_v$. The estimates could be linear or non-linear functions (g_a, g_v) depending on the specific interaction prior. This model can describe complete fusion, partial integration, or segregation of cues. **(C)** Causal inference model. The latent variable C determines the causal structure that generates the cues and mediates cue integration: cues are integrated if they have a common cause ($C = 1$) and processed separately if they have independent causes ($C = 2$). The model infers the probability of the unknown causal structure p(C |X_v, X_a) and weights the estimates $\hat{S}_a$ and $\hat{S}_v$ accordingly using some decision strategy (Wozny et al., 2010). The estimates are nonlinear combinations of the cues and usually require Monte Carlo simulation to obtain (Körding et al., 2007). This model can be recast as the coupling prior model **(B)** by integrating out the latent variable C, in which case it will no longer explicitly represent the causal structure.

minimum-variance estimate typically takes into account the full correlation matrices of the noise (Doclo et al., 2015).

Recent psychophysical research has suggested that the MLE-type complete fusion is not a general property of human multisensory perception (e.g., Battaglia et al., 2003; Arnold et al., 2019; Meijer et al., 2019). To capture the full spectrum of cue interaction spanning from complete fusion to partial integration to segregation, extensions of the forced fusion model have been proposed. Among them, the *coupling prior model* (**Figure 1B**), also known as the interaction prior model, extends the forced fusion model (**Figure 1A**) by adding a joint prior distribution to represent the correlation or co-occurrence statistics between the cues (Shams et al., 2005; Rowland et al., 2007; Ernst, 2012; Parise et al., 2014). For example, in a speech recognition task with auditory and visual cues, a coupling prior model could use a bivariate prior $p(S_a,\ S_v)$ to describe the joint probability distribution for the auditory (S_a) and visual (S_v) representations of the stimulus attribute (e.g., syllables). The coupling prior can be conveniently modeled using a 2D Gaussian $p(S_a,\ S_v) = N_{Sa,\ Sv}(\bar{s},\ \Sigma)$, with the mean $\bar{s}$ being the expected stimulus value, and the covariance matrix Σ consisting of variances along the principle axes (e.g., Ernst, 2007). The $p(S_a,\ S_v)$ distribution is sharper if the AV coupling is relatively constant (due to statistical regularities in the environment or acquired through adaptation or learning). The forced fusion model is a special case of the coupling prior model where $p(S_a, S_v) = 0$ for all $S_a \neq\ S_v$. Another method for characterizing the coupling prior is to use a GMM to represent the correlated and the uncorrelated components (e.g., Roach et al., 2006; Sato et al., 2007); the resulting mixture estimator is more general and robust than MLE.

The coupling prior model for cue integration is analogous to a GMM for AVSR, where the AV coherence (i.e., dependency between the auditory and visual modalities) is expressed as a joint AV PDF (Rivet et al., 2014). It can be viewed as loosely similar to the basic concept of coupled HMMs for AVSR, too. However, unlike coupled HMMs, the coupling prior model is not dynamic and does not describe time-variant signals. Moreover, the coupling prior model explicitly constrains the joint prior distribution of the cues, whereas coupled HMMs implicitly learn the hidden states that generate the cues.

SOURCE SEPARATION AND CAUSAL INFERENCE

In machine SP, the most common scenario of source separation is *blind source separation* (BSS): separating two or more source signals given mixture observations (Jutten and Herault, 1991; Castella et al., 2010). A fundamental challenge in BSS is the *label permutation problem*: to track which speech signal belongs to which speaker/source (Hershey et al., 2016). To achieve this, a model needs to jointly solve two problems: isolating a single speech signal from a dynamic mixture of sounds from multiple speakers and the background noise, and assigning the speech signal to the corresponding speaker (Ephrat et al., 2018). A Bayesian approach to solve BSS is applying GMMs and HMMs that either constrain or learn the unobservable source structure underlying the mixture signals (Roweis, 2001, 2003; Hershey and Casey, 2002; Yilmaz and Rickard, 2004). Inspired by human perception, recent machine SP models have been exploiting the intrinsic AV coherence to improve BSS performance (Rivet et al., 2014). Full joint AV models based on maximizing the AV likelihood can successfully extract source signals from underdetermined mixtures (Sodoyer et al., 2002). However, such models are limited to *instantaneous* mixtures, where multiple source signals contribute to the mixtures without delay at a given time point. Similarly in human perception, most existing mixture models for cue integration consider only instantaneous mixtures (e.g., Magnotti and Beauchamp, 2017). If multiple source signals contribute to the mixtures with different levels of delay—known as *convolutive* mixtures—alternative techniques are required to resolve the added ambiguities in BSS (e.g., Rivet et al., 2007; Liu et al., 2012. For a review, see Rivet et al., 2014).

In natural environments, the structure of the source(s) giving rise to the signals is often ambiguous or unobservable; therefore, to properly associate a signal with its source, the observer needs to infer cause-effect relationships based on the noisy data. This is an example of the so-called *inverse problem* in information processing: inferring the cause given the effect (Ghahramani, 2015). Humans are remarkably apt at solving this problem, being able to focus on a target speaker while filtering out interfering sounds and background noise, as exemplified by the well-known cocktail party effect (Cherry, 1953). However, the causal inference problem is challenging for machine SP, especially in AVSR, as it is difficult to determine which signals in the mixture data came from the same source and thus should be fused.

Machine SP could draw inspiration from the *causal inference model* in human psychophysics (**Figure 1C**), which explicitly characterizes the hidden causal structure of the source signal(s) (Körding et al., 2007; Shams and Beierholm, 2010; Magnotti and Beauchamp, 2017). This model proposes that humans estimate the hidden causal structure based on statistical regularities of the environment and use this estimate to arbitrate between grouping or segregating sensory cues (Noppeney and Lee, 2018). The basic structure of this model has two hierarchies. In the higher hierarchy is a binary latent variable representing whether the multiple cues share a common cause, denoted as C (short for "cause"). $C = 1$ means the cues have a common cause, and $C = 2$ means the cues have two separate causes. The *a-priori* belief for C is the *causal prior*, and it influences whether and to which degree cues are integrated: cues are integrated only if they have a common cause, in which case the model is equivalent to a forced-fusion MLE model (**Figure 1A**); in contrast, the cues are processed separately if they originate from different causes. The causal structure is unknown, so the model needs to infer C by combining bottom-up sensory data with top-town causal priors and calculating the posterior $p\,(C|X_a,\ X_v)$ for different C values. The model additionally computes the PDF for the task-relevant estimate $p\left(\hat{S}\middle|X_a,\ X_v,\ C\right)$ under the assumption of common or separate causes, respectively. A final estimate for the stimulus attribute is obtained by combining these estimates according to some decision strategy. For example, if a model-averaging decision strategy is applied, which is based on the use of MMSE, then the resulting final estimate is the weighted average of the estimates obtained under $C = 1$ and $C = 2$, respectively, with the weights being the corresponding posterior probabilities for $C = 1$ and $C = 2$ (Körding et al., 2007; Wozny et al., 2010).

SUMMARY AND OUTLOOK

Here we reviewed a selection of probabilistic models of audio- and AV-processing applied in machine SP and in human perception, focusing on speech enhancement, speech recognition, cue integration, and causal inference (**Table 1**). In their cores, these models are stimulus-response functions: they describe a probability distribution of responses given a stimulus and parameters, and the parameters can be estimated from experimental data or machine learning methods. Basic probabilistic models are often linear filters with Gaussian PDFs (e.g., Wiener filter, classic STRF), which can be extended with nonlinear, adaptive, and/or top-down features (e.g., super-Gaussian prior, gain control, selective attention). In addition,

TABLE 1 | An overview of selective probabilistic models of audio- and audio-visual (AV) processing in machines and human perception.

	Problem	Model	Main features and advantages	Limitations
		Noise reduction and speech enhancement		
Machine speech enhancement	Estimation of speech coefficients	Wiener filter with simple Gaussian PDFs	Linear, low computational cost, easy to implement	Gaussian PDFs not appropriate for modeling speech Fourier coefficients. Super-Gaussian is better
		MVDR beamformer	Suitable for multi-channel noise reduction	
		GMM	Dynamics of speech and noise captured by states of a mixture model. Mixture estimator	Typically restricted to a small number of classes; limited robustness in reverberant conditions
		HMM	Improves modeling of temporal behavior by including state transitions. Mixture estimator	Strong restrictive assumptions, intolerant to state asynchrony in AV combined streams, sensitive to initial parameter values
Auditory neural processing	Maintaining robust neuronal representation of relevant sounds	Spectrotemporal receptive field (STRF)	Computational simplicity, analytic tractability, interpretability	Does not capture the highly nonlinear and dynamic features of auditory neurons
		Audio-visual (AV) integration and speech recognition		
Machine ASR	ASR	GMM/HMM	Captures the dynamics of speech	Other modalities cannot be easily included
	AVSR	Coupled HMM, factorial HMM	Improves AV fusion over conventional HMM for AVSR	
Human AV integration	Optimal AV cue combination	Forced fusion (MLE) model	Reliability-weighting, minimum-variance unbiased estimator	Complete fusion only; does not account for cue coherence or causal structure
	Accounting for AV correlation	Coupling prior model, can use GMM	Joint AV prior distribution. Can capture the full range of AV integration	Cannot infer causal relationships
		Source separation and causal inference		
Machine source separation	Source separation, label permutation problem	Blind source separation techniques with GMM, HMM, etc.	Does not need a-priori information about causal structure; works for instantaneous mixtures and convolutive mixtures	
Human AV integration	Causal inference	Causal inference model	Explicitly represents the underlying causal structure; more general than forced-fusion and coupling prior models	Can be computationally expensive

the use of mixture models (e.g., GMM, HMM) simultaneously accounts for multiple possible states and permits more robust parameter estimation. Furthermore, basic probabilistic models can be adapted to characterize multiple input channels or streams (e.g., MVDR beamformer). If multiple inputs are combined (e.g., cue integration, AVSR), fusion models with reliability-based weighting and MLE are typically applied. However, forced fusion is not always appropriate. Therefore, to capture the large spectrum of input interactions, some models incorporate the correlation between the inputs (e.g., coupling prior model, coupled or factorial HMM) instead of assuming fusion. Moreover, causal inference models estimate the hidden source or causal structure of the inputs, by factoring in causality which is important for determining input integration or source separation. More advanced models, such as those in machine learning, are beyond the scope of this mini review. In short, this brief tutorial linked the analogous counterparts among probabilistic models developed in artificial and natural systems and identified the closest points of potential overlap between these models.

AUTHOR CONTRIBUTIONS

All authors: conceptualization, review and editing, and approval of final version. LL and RR: literature research and analysis and manuscript drafting.

FUNDING

This work was supported by the German Research Foundation (DFG) grant TRR 169/A1 to PB and BR, the DFG TRR 169/A6 to TG, and Die Hamburger Behörde für Wissenschaft, Forschung und Gleichstellung (BWfG) (City of Hamburg, Department of Science, Research, and Equality) ahoi.digital project "Adaptive crossmodal sensor data acquisition" to RR.

REFERENCES

Abdelaziz, A. H., Zeiler, S., and Kolossa, D. (2015). Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition. *IEEE ACM Trans. Audio Speech Lang. Process.* 23, 863–876. doi: 10.1109/TASLP.2015.2409785

Adjoudani, A., and Benoît, C. (1996). "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines, Models, Systems and Applications of NATO ASI Series F: Computer and Systems Sciences,* Vol. 150, eds D. G. Stork and M. E. Hennecke (Berlin: Springer-Verlag), 461–471.

Ahrens, M. B., Linden, J. F., and Sahani, M. (2008). Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *J. Neurosci.* 28, 1929–1942. doi: 10.1523/JNEUROSCI.3377-07.2008

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Arnold, D. H., Petrie, K., Murray, C., and Johnston, A. (2019). Suboptimal human multisensory cue combination. *Sci. Rep*. 9:5155. doi: 10.1038/s41598-018-37888-7

Balan, R., and Rosca, J. (2002). "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings* (Rosslyn, VA: IEEE), 209–213.

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J. Opt. Soc. Am. A* 20:1391. doi: 10.1364/JOSAA.20.001391

Brand, M., Oliver, N., and Pentland, A. (1997). "Coupled hidden markov models for complex action recognition," in *Proceeding IEEE International Conference on Computer Vision and Pattern Recognition* (San Juan), 994–999.

Burshtein, D., and Gannot, S. (2002). Speech enhancement using a mixture-maximum model. *IEEE Trans. Speech Audio Process.* 10, 341–351. doi: 10.1109/TSA.2002.803420

Calabrese, A., Schumacher, J. W., Schneider, D. M., Paninski, L., and Woolley, S. M. N. (2011). A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PLoS ONE* 6:e16104. doi: 10.1371/journal.pone.0016104

Castella, M., Chevreuil, A., and Pesquet, J.-C. (2010). "Convolutive mixtures," in *Handbook of Blind Source Separation*, eds P. Common and C. Jutten (New York, NY: Elsevier, Academic Press), 281–324.

Chazan, S. E., Goldberger, J., and Gannot, S. (2016). A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier. *IEEE ACM Trans. Audio Speech Lang. Process.* 24, 2516–2530. doi: 10.1109/TASLP.2016.2618007

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Netw. Comput. Neural Syst*. 12, 199–213. doi: 10.1080/713663221

Colonius, H., and Diederich, A. (2018). Formal models and quantitative measures of multisensory integration: a selective overview. *Eur. J. Neurosci.* 51, 1161–1178. doi: 10.1111/ejn.13813

David, S. V. (2018). Incorporating behavioral and sensory context into spectro-temporal models of auditory encoding. *Heart Res*. 360, 107–123. doi: 10.1016/j.heares.2017.12.021

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2144–2149. doi: 10.1073/pnas.1117717109

Deng, L., and Li, X. (2013). Machine learning paradigms for speech recognition: an overview. *IEEE Trans. Audio Speech Lang. Process.* 21, 1060–1089. doi: 10.1109/TASL.2013.2244083

Doclo, S., Kellermann, W., Makino, S., and Nordholm, S. E. (2015). Multichannel signal enhancement algorithms for assisted listening devices: exploiting spatial diversity using multiple microphones. *IEEE Signal Process. Mag*. 32, 18–30. doi: 10.1109/MSP.2014.2366780

Ephraim, Y. (1992). A bayesian estimation approach for speech enhancement using hidden markov models. *IEEE Trans. Signal Process.* 40, 725–735. doi: 10.1109/78.127947

Ephraim, Y., and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust*. 32, 1109–1121. doi: 10.1109/TASSP.1984.1164453

Ephraim, Y., and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust.* 33, 443–445. doi: 10.1109/TASSP.1985.1164550

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., et al. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph* 37:109. doi: 10.1145/3197517.3201357

Ernst, M. (2007). Learning to integrate arbitrary signals from vision and touch. *J. Vis*. 7:7. doi: 10.1167/7.5.7

Ernst, M. O. (2012). "Optimal multisensory integration: assumptions and limits," in *The New Handbook of Multisensory Processes*, ed B. Stein (Cambridge, MA: MIT Press), 527–544.

Ernst, M. O., and Bülthof, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169. doi: 10.1016/j.tics.2004.02.002

Fetsch, C. R., DeAngelis, G. C., and Angelaki, D. E. (2013). Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Neurosci.* 14, 429–442. doi: 10.1038/nrn3503

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci*. 6, 1216–1223. doi: 10.1038/nn1141

Gerkmann, T., and Hendriks, R. C. (2012). Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio Speech Lang. Process.* 20, 1383–1393. doi: 10.1109/TASL.2011.2180896

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature* 521, 452–459. doi: 10.1038/nature14541

Ghahramani, Z., and Jordan, M. I. (1997). Factorial hidden markov models. *Mach. Learn*. 29, 245–273. doi: 10.1023/A:1007425814087

Hendriks, R. C., Gerkmann, T., and Jensen, J. (2013). "DFT-domain based single-microphone noise reduction for speech enhancement - a survey of the state of the art," in *Synthesis Lectures on Speech and Audio Processing* (San Rafael, CA: Morgan & Claypool Publishers), 1–80.

Hennecke, M. E., Stork, D. G., and Prasad, K. V. (1996). "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines, Models, Systems and Applications, Volume 150 of NATO ASI Series F: Computer and Systems Sciences*, eds D. G. Stork and M. E. Hennecke (Berlin: Springer-Verlag), 331–349.

Hershey, J., Attias, H., Jojic, N., and Kristjansson, T. (2004). "Audio-visual graphical models for speech processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 5 (Montreal, QC), 649–652.

Hershey, J. R., and Casey, M. (2002). "Audio-visual sound separation via hidden markov models," in *Advances in Neural Information Processing Systems (NIPS),* eds T. G. Dietterich, S. Becker, and Z. Ghahramani (Vancouver, BC: MIT Press), 1173–1180.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). "Deep clustering: discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai), 31–35.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intel. Ligence* 22, 4–37. doi: 10.1109/34.824819

Jutten, C., and Herault, J. (1991). Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. *Signal Process.* 24, 1–10. doi: 10.1016/0165-1684(91)90079-X

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing - Volume 1: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall.

King, A. J., Teki, S., and Willmore, B. D. B. (2018). Recent advances in understanding the auditory cortex. *F1000Research* 7:1555. doi: 10.12688/f1000research.15580.1

Kolossa, D., and Häb-Umbach, R. (2011). *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, *1st Edn*. Berlin; Heidelberg: Springer.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE* 2:e943. doi: 10.1371/journal.pone.0000943

Krawczyk-Becker, M., and Gerkmann, T. (2016). Fundamental frequency informed speech enhancement in a flexible statistical framework. *IEEE ACM Trans. Audio Speech Lang. Proc.* 24, 940–951. doi: 10.1109/TASLP.2016.2533867

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav Brain Sci.* 40:e253. doi: 10.1017/S0140525X16001837

Lee, T., and Theunissen, F. (2015). A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 471:20150309. doi: 10.1098/rspa.2015.0309

Liu, Q., Wang, W., and Jackson, P. (2012). Use of bimodal coherence to resolve per- mutation problem in convolutive BSS. *Signal Process.* 92, 1916–1927. doi: 10.1016/j.sigpro.2011.11.007

Lohse, M., Bajo, V. M., King, A. J., and Willmore, B. D. B. (2020). Neural circuits underlying auditory contrast gain control and their perceptual implications. *Nat. Commun.* 11, 324. doi: 10.1038/s41467-019-14163-5

Lotter, T., and Vary, P. (2005). Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP J. Adv. Signal Process.* 2005:354850. doi: 10.1155/ASP.2005.1110

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends Cogn. Sci.* 16, 511–518. doi: 10.1016/j.tics.2012.08.010

Magnotti, J. F., and Beauchamp, M. S. (2017). A causal inference model explains perception of the mcgurk effect and other incongruent audiovisual speech. *PLoS Comput. Biol.* 13:e1005229. doi: 10.1371/journal.pcbi.1005229

Maloney, L. T. (2002). "Statistical theory and biological vision," in *Perception and the Physical World: Psychologocal and Philosophical Issues in Perception*, eds D. Heyer and R. Mausfeld (Neywork, NY: Wiley), 145–189.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* 9, 504–512. doi: 10.1109/89.928915

Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.* 13, 845–856. doi: 10.1109/TSA.2005.851927

Meijer, D., Veselič, S., Calafiore, C., and Noppeney, U. (2019). Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex* 119, 74–88. doi: 10.1016/j.cortex.2019.03.026

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1–6. doi: 10.1073/pnas.1318017111

Meutzner, H., Ma, N., Nickel, R., Schymura, C., and Kolossa, D. (2017). "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Proceeding ICASSP* (New Orleans, LA: IEEE).

Meyer, A. F., Williamson, R. S., Linden, J. F., and Sahani, M. (2017). Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. *Front. Syst. Neurosci.* 10:109. doi: 10.3389/fnsys.2016.00109

Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. (2002). Dynamic bayesian networks for audio-visual speech recognition. *EURASIP J. Adv. Signal Process.* 2002, 1274–1288. doi: 10.1155/S1110865702206083

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., et al. (2000). "Audio Visual Speech Recognition," in *Workshop 2000 Final Report*. (Baltimore, MD). Available online at: https://pdfs.semanticscholar.org/c099/a6dc8b9f731bbc9a29a920233257c13d9b00.pdf

Noppeney, U., and Lee, H. L. (2018). Causal inference and temporal predictions in audiovisual perception of speech and music. *Ann. N. Y. Acad. Sci.* 1423, 102–116. doi: 10.1111/nyas.13615

Padmanabhan, J., and Premkumar, M. J. J. (2015). Machine learning in automatic speech recognition: a survey. *IETE Tech. Rev.* 32, 240–251. doi: 10.1080/02564602.2015.1010611

Paninski, L. (2003). Convergence properties of some spike-triggered analysis techniques. *Network: Comput Neural Syst.* 14, 437–464. doi: 10.1088/0954-898X_14_3_304

Parise, C. V., Knorre, K., and Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6104–6108. doi: 10.1073/pnas.1322705111

Porter, J., and Boll, S. (1984). "Optimal estimators for spectral restoration of noisy speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (San Diego, CA), 53–56.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* 91, 1306–1325. doi: 10.1109/JPROC.2003.817150

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains : knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Rabinowitz, N. C., Willmore, B. D. B., King, A. J., and Schnupp, J. W. H. (2013). Constructing noise-invariant representations of sound in the auditory pathway. *PLoS Biol* 11:e1001710. doi: 10.1371/journal.pbio.1001710

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., and King, A. J. (2012). Spectrotemporal contrast kernels for neurons in primary auditory cortex. *J. Neurosci.* 32, 11271–11284. doi: 10.1523/JNEUROSCI.1715-12.2012

Rao, R. P., Olshausen, B. A., Lewicki, M. S., Jordan, M. I., and Dietterich, T. G. (2002). *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.

Rehr, R., and Gerkmann, T. (2018). On the importance of super-gaussian speech priors for machine-learning based speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 357–366. doi: 10.1109/TASLP.2017.2778151

Rehr, R., and Gerkmann, T. (2019). "An analysis of noise-aware features in combination with the size and diversity of training data for DNN-based speech enhancement," in *IEEE International Conference Acoustics Speech Signal Process (ICASSP)* (Brighton).

Rivet, B., Girin, L., Servi?re, C., Pham, D.-T., and Jutten, C. (2007). "Using a visual voice activity detector to regularize the permutations in blind source separation of convolutive speech mixtures," in *Proceeding International Conference Digital Signal Processing (DSP)* (Cardiff), 223–226.

Rivet, B., Wang, W., Naqvi, S. M., and Chambers, J. A. (2014). Audiovisual speech source separation: an overview of key methodologies. *IEEE Signal Process. Mag.* 31, 125–134. doi: 10.1109/MSP.2013.2296173

Roach, N. W., Heron, J., and Mcgraw, P. V. (2006). Resolving multisensory conflict : a strategy for balancing the costs and benefits of audio-visual integration. *Proc. R. Soc. B Biol. Sci.* 273, 2159–2168. doi: 10.1098/rspb.2006.3578

Rohde, M., van Dam, L. C. J., and Ernst, M. O. (2015). Statistically optimal multisensory cue integration?: A practical tutorial. *Multisens. Res.* 1–39. doi: 10.1163/22134808-00002510

Roweis, S. T. (2001). "One microphone source separation," in *Advances in Neural Information Processing Systems 13*, eds T. K. Leen, T. G. Dietterich, and V. Tresp (Vancouver, BC: MIT Press), 793–799.

Roweis, S. T. (2003). "Factorial models and refiltering for speech separation and denoising," in *Eurospeech* (Geneva).

Rowland, B., Stanford, T., and Stein, B. (2007). A Bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Exp. Brain Res.* 180, 153–161. doi: 10.1007/s00221-006-0847-2

Sato, Y., Toyoizumi, T., and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect : identification of common sources. *Neural Comput.* 19, 3335–3355. doi: 10.1162/neco.2007.19.12.3335

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition* 93, B69–B78. doi: 10.1016/j.cognition.2004.01.006

Shams, L., and Beierholm, U. R. (2010). Causal inference in perception. *Trends Cogn. Sci.* 14, 425–432. doi: 10.1016/j.tics.2010.07.001

Shams, L., Ma, W. J., and Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport* 16, 1923–1927. doi: 10.1097/01.wnr.0000187634.68504.bb

Sharpee, T., Rust, N. C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.* 16, 223–250. doi: 10.1162/089976604322742010

Sodoyer, D., Schwartz, J.-L., Girin, L., Klinkisch, J., and Jutten, C. (2002). Separation of audio-visual speech sources: a new approach exploiting the audio-visual coherence of speech stimuli. *EURASIP J. Adv. Signal Process.* 2002, 1165–1173. doi: 10.1155/S1110865702207015

Theis, L., Chagas, A. M., Arnstein, D., Schwarz, C., and Bethge, M. (2013). Beyond GLMs: a generative mixture modeling approach to neural system identification. *PLoS Comput. Biol* 9:e1003356. doi: 10.1371/journal.pcbi.1003356

Ursino, M., Cuppini, C., and Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Netw.* 60, 141–165. doi: 10.1016/j.neunet.2014.08.003

Willmore, B. D. B., Cooke, J. E., and King, A. J. (2014). Hearing in noisy environments: noise invariance and contrast gain control. *J. Physiol.* 592, 3371–3381. doi: 10.1113/jphysiol.2014.274886

Willmore, B. D. B., Schoppe, O., King, A. J., Schnupp, J. W. H., and Harper, N. S. (2016). Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. *J. Neurosci.* 36, 280–289. doi: 10.1523/JNEUROSCI.2441-15.2016

Wozny, D. R., Beierholm, U. R., and Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Comput. Biol.* 6:e1000871. doi: 10.1371/journal.pcbi.1000871

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci* 19, 356–365. doi: 10.1038/nn.4244

Yilmaz, O., and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.* 52, 1830–1847. doi: 10.1109/TSP.2004.828896

Yuille, A. L., and Bülthoff, H. H. (1996). "Bayesian decision theory and psychophysics," in *Perception as Bayesian Inference*, eds D. C. Knill and W. Richards (Cambridge: Cambridge University Press), 123–162.

Yumoto, E., Gould, W. J., and Baer, T. (1982). Harmonic to noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Am.* 71, 1544–1550. doi: 10.1121/1.387808

Zhao, D. Y., and Kleijn, W. B. (2007). HMM-based gain modeling for enhancement of speech in noise. *IEEE Trans. Audio Speech Lang. Process.* 15, 882–892. doi: 10.1109/TASL.2006.885256

Zhao, L., and Zhaoping, L. (2011). Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles. *PLoS Comput. Biol* 7:e1002123. doi: 10.1371/journal.pcbi.1002123

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.