






Dong Yu , Yifan Gong, Michael Alan Picheny, Bhuvana Ramabhadran , Dilek Hakkani-Tür, Rohit Prasad, Heiga Zen, Jan Skoglund , Jan "Honza" Černocký , Lukáš Burget , and Abdelrahman Mohamed

Twenty-Five Years of Evolution in Speech and Language Processing



©SHUTTERSTOCK.COM/TRIFF

In this article, we summarize the evolution of speech and language processing (SLP) in the past 25 years. We first provide a snapshot of popular research topics and the associated state of the art (SOTA) in various subfields of SLP 25 years ago, and then highlight the shift in research topics over the years. We describe the major breakthroughs in each of the subfields and the main driving forces that led us to the SOTA today. Societal impacts and potential future directions are also discussed.

Introduction

The year 2023 marks the 75th anniversary of the IEEE Signal Processing Society (SPS). Technologies have been significantly advanced in these 75 years, and society has been greatly impacted by these advances. For example, the mobile Internet has greatly changed people's lifestyles. Researchers and practitioners in signal processing have contributed their share to these progresses.

In this article, we concentrate on the field of SLP, which is the scope covered by the IEEE Speech and Language Processing Technical Committee (SLTC), and summarize the major technological developments in the field and the key societal impacts caused by these advances in the past 25 years.

As part of the SPS, the SLTC serves, promotes, and influences all the technical areas of SLP, including automatic speech recognition (ASR), speech synthesis [often referred to as *text to speech (TTS)*], speaker recognition (SPR) and diarization, language identification (LID), speech enhancement, speech coding, speech perception, language understanding, and dialog systems.

The SLTC can trace its roots back to the Institute of Radio Engineers Audio Group, founded in 1947. In 1969, this audio group established the Speech Processing and Sensory Aids Technical Committee. "Sensory Aids" was dropped from the name in the early 1970s. For more than 30 years, it remained the Speech Processing Technical Committee. In 2006, its scope was expanded, and its name was officially changed to the SLTC.

Today, more than 50 years after the formation of the SLTC and 16 years after the recent name change, the field has been

significantly expanded and greatly reshaped by new thoughts and techniques. In fact, we have observed rapid progress in SLP in the past decade, largely driven by deep learning, big data, high-performance computation, and application demands. For example, ASR accuracy has surpassed the adoption threshold in the closed-talk setup where the microphone is close to the mouth. TTS can now generate natural-sounding speech that is hard to distinguish from human speech [1]. The performance of natural language processing tasks has been greatly improved with huge, pretrained language models (LMs).

The remainder of the article is organized as follows. In the “Status of the Field in 1998” section, we provide an overview of the field and summarize the main knowledge and key issues of each major subfield 25 years ago. In the “Main Driving Forces Over the Last Decades” section, we describe the main driving forces that have reshaped the field and caused a paradigm change in the past decade. In the “Major Technical Breakthroughs in Each Subfield” section, we summarize major breakthroughs and the current SOTA in each subfield. In the “Conclusion” section, we conclude the article with comments on the societal impact, and perspectives on future developments in the domain.

Status of the field in 1998

SLP has been an active research area since the 1950s. By 1998, the field had already made great leaps. The many key technologies that we know of today were developed then. In this section, we provide an overview of the field, summarize the main knowledge, and point out the key technical obstacles at that time.

Overview of the field

IEEE played an important role in pushing the SOTA of the field around 1998. *IEEE Transactions on Speech and Audio Processing* and *ICASSP* were the flagship journal and conference, respectively. In 1997, the SLTC extended the scope of the IEEE Automatic Speech Recognition Workshop and renamed it the *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Almost all of the popular subfields we study today had been extensively studied by 1998. The proceedings of *ICASSP* indicate that in 1998, the popular topics were ASR, speech enhancement, speech coding and perception, speech synthesis and analysis, speaker and LID, and speech-to-speech translation. The topic of spoken language understanding (SLU) was better covered in *ASRU*.

Speech coding

The task of speech coding is to compress speech signals for efficient storage and transmission. The major speech coding event around 1998 was the launch of the mixed-excitation linear prediction (MELP) codec [2], which is based on the linear prediction coder (LPC) but with five additional features, presented in 1997 as the winning candidate of the U.S. Department of Defense contest to select a new federal standard for narrow-band (8-kHz sampling frequency) voice coding at

2.4 kbps, replacing the previous LPC10 codec from 1984. In a way, this contest ended a decade-long golden age of speech coding as the growth of digital mobile telephony applications in the late 1980s through the 1990s required rapid development in the field.

At the time, speech coding algorithms could be categorized into two classes: model-based parametric codecs (also called *source codecs* or *vocoders*) such as MELP, operating mainly at rates up to 5 kbps, and waveform-matching codecs at rates above 5 kbps. Model-based parametric codecs usually consider the source-filter model of speech production and preserve only the spectral properties of speech. By sending only the source/excitation type and the filter parameters, model-based parametric codecs can achieve very low bit rates. Waveform-matching codecs aim to reproduce the speech waveform as faithfully as possible. They usually do not rely on the prior knowledge that might have created the audio.

Even though the waveform-matching, linear-predictive codecs based on code-excited linear prediction (CELP) [3] employed a model separating the speech signal into an excitation signal driving a linear synthesis filter, the analysis by synthesis offered monotonically increased fidelity with an increased bit rate. Basically, all mobile telephony standards are based on the CELP methodology. At bit rates below 5 kbps, waveform matching performs poorly, and better quality is achieved by model-based parametric coding with efficient quantization of extracted speech features and vocoder synthesis. Parametric codecs have a tendency to produce speech that sounds a bit unnatural and robotic. The speech quality is limited by the model; hence, quality will stop improving after certain rates. The major application for parametric coders was secure voice communication, where bandwidth was limited and speech intelligibility was more important than high-fidelity quality.

After 1998, research interest in narrow-band, low-rate speech coding subsided. For example, there was a subsequent International Telecommunications Union telecommunication standardization sector standardization effort for a new 4-kbps speech codec a few years later. This effort was essentially abandoned due to irrelevancy after the emergence of Voice over Internet Protocol (IP), and future mobile communication standards offered higher bandwidths and encouraged speech compression efforts toward higher bit rates and sampling rates.

ASR

The task of ASR is to convert the speech audio sequence into the corresponding text sequence. The modern field of speech recognition, as a whole, had its origins in information theory as far back as the early 1970s. For a fascinating view of how this approach took hold, the reader is referred to [4]. By 1998, data-driven approaches employing complex statistical models had gained broad acceptance in the community. Indeed, ASR was already being viewed as a largely solved problem, with a large push being made toward the commercialization of the technology.

Around 1998, a typical speech recognition system comprised the following components:

- *Feature vector extractor*: mel-frequency cepstral coefficients (MFCCs) or perceptual linear prediction coefficients
- *Acoustic model (AM)*: a context-dependent, phonetic hidden Markov model (HMM) using Gaussian mixture models (GMMs) to model feature vector probabilities
- *LM*: a backoff-based n -gram LM, sometimes class-based LM
- *Hypothesis search*: a Viterbi decoder, beam search; single or sometimes multipass
- *Adaptation mechanisms*: an AM adaptation with maximum-likelihood linear regression or maximum a posteriori (MAP). LM adaptation using cache-based interpolation with an add-word mechanism to handle out-of-vocabulary words.

It should be noted that neural network (NN) technology had achieved some limited success by 1998, but that its performance was not good enough to replace GMMs, much less HMMs. The early systems were all trained using the maximum-likelihood criterion. By 1998, discriminative training criteria already existed, but performance gains were small and expensive to achieve. Back in 1998, “large” systems were trained on a few hundred hours of speech.

TTS synthesis

The goal of TTS is to render naturally sounding speech given an arbitrary text. Two data-driven, corpus-based speech synthesis approaches were proposed in the 1990s: an exemplar-based, unit-selection approach with which speech is synthesized by concatenating scaled pronunciation-unit samples from a corpus, and a model-based, generative approach.

Shortly after the proposal of the first data-driven unit-selection TTS [5] in 1992, an HMM-based, trainable unit-selection speech synthesis was proposed [6], where decision tree-clustered, context-dependent HMMs were used for unit segmentation and cost functions. The formulation and trainable framework of unit selection made it popular in R&D for the next two decades.

At the same time, the first paper toward generative TTS was proposed in 1995 [7], where probabilities of acoustic features (vocoder parameters) given linguistic features (context-dependent phonemes) were modeled and generated using HMMs. The generative TTS’s flexibility to change its voice characteristics was demonstrated in [8]. The generative approach was still incomplete in 1998 as it lacked prosody modeling and generation. Prosody refers to the duration, intonation, and intensity patterns of speech associated with the sequence of syllables, words, and phrases. Without proper prosody modeling and generation, long sentences will sound unnatural.

Although the unit-selection approach could synthesize naturally sounding speech for in-domain texts (those covered well by the corpus), due to data sparsity, its quality could degrade significantly for out-of-domain texts with discontinuities in the generated speech caused by uncovered units. Furthermore, having multiple speakers, emotions, and speaking styles was difficult as it required an ample number of recordings with these characteristics. On the other hand, the generative approach had already demonstrated a way to change its voice characteristics. However,

the naturalness of synthesized speech was limited by the quality of vocoders.

SPR, identification, and diarization

The task of SPR infers speaker identity from the speech signal. The most straightforward task is speaker verification, which aims to determine whether two recordings were spoken by the same speaker or different ones. A range of other tasks can be derived from speaker verification, such as speaker identification (closed or open set), speaker tracking (determining speaker trajectories), and speaker search (determining from where a specific voice is speaking). Two basic settings are text dependent (the speaker needs to provide a predetermined key phrase) and text independent. Speaker diarization is a derived, more complicated task as it aims to segment a recording into regions spoken by one speaker and generate speaker labels (such as “A,” “B,” and so on) consistently over the recording.

The status around 1998 is covered in an excellent tutorial by Campbell [9]. A typical SPR system is a statistical model that contained feature extraction, pattern matching, and decision (see Figure 1). As in related fields, all the possibilities of features (LPC, MFCC, line spectral pairs, and so on) were investigated, and several matching techniques (Gaussian modeling, distance computation, and dynamic time warping) competed. R&D usually contained feature selection, testing, and the fusion of several matching techniques. Several researchers experimented with NNs, but without much success. The National Institute of Standards and Technology’s (NIST’s) Speaker Recognition Evaluation series started in 1996 and has since become a platform to evaluate SPR technology.

In speaker diarization, a typical system in 1998 already contained similar components as the current ones (excluding the end-to-end ones): segmentation and automatic clustering of segments. Kullback–Leibler (KL) distance was widely used. For the segmentation, a sequence of cepstral features was extracted from the input speech and split into 2-s windows. A segment boundary was detected if the KL distance between Gaussian distributions estimated for the neighboring windows was above a threshold. Similarly, it was also used for agglomerative clustering of the resulting segments, i.e., initially treating each segment as a cluster and then gradually merging clusters. The 1996 DARPA Hub 4 Broadcast News Evaluation was a popular task used to evaluate diarization.

LID

LID, also termed *spoken language recognition* or just *language recognition (LR)*, aims to determine the language in a particular speech segment. Engineers usually depend on linguists and politicians (https://en.wikipedia.org/wiki/A_language_is_a_dialect_with_an_army_and_navy) to answer the “Language or dialect?” question and consider every class labeled with the same label as a language.

Around 1998, two standard LID approaches were defined [10]: acoustic, aiming at the classification of a sequence of acoustic feature vectors into a class, and phonotactic, first tokenizing the input sequence into discrete units (phones)

using one or several phone recognizers, and then performing language (phonotactic) modeling. Each class (language) was modeled by a GMM in the case of the acoustic approach, and by an n -gram LM in the case of phonotactic systems. The latter obviously depended on the accuracy of phone recognition. Similar to SPR, LID was driven by the NIST Speaker Recognition Evaluation series started in 1996.

Speech enhancement and separation

Real-world speech signals are often contaminated by interfering speakers, noises, and reverberation. The task of speech enhancement and separation, which aims at extracting clean speech signals from a mixture, is thus very important for both human-to-human and human-machine communication. Conventionally, people refer to *speech separation* as the problem of segregating one or more target speakers from other interfering speakers and/or noises, and *speech enhancement* as the problem of removing noises and/or reverberation from noisy speech.

The dominant techniques for speech enhancement were purely signal processing-based in 1998 [11]. Under this framework, enhancement of noisy speech signals is essentially a problem of estimating a clean speech signal from a given sample function of the noisy signal by minimizing the expected value of some distortion measure between the clean and estimated signals. Although these techniques (e.g., the Wiener filter) differ in the statistical models (e.g., Gaussian and hidden Markov processes) assumed, the distortion metric (e.g., minimum mean-square error) used, the domain (e.g., time, frequency, and magnitude domain) in which the enhancement is carried out, and the way the noise and speech statistics are estimated (e.g., minimum statistics-based noise estimator), they often assume that speech and noise follow statistically independent processes.

Speech separation is usually considered a more difficult problem because the target speech and the interfering speech share very similar characteristics. At that time, the main approach to speech separation, which focused on blind source separation, was independent component analysis, which aims at recovering a set of maximally independent sources from the observed mixtures without knowing the source signals or the mixing parameters. Also in that time period, the perceptual principles of human auditory scene analysis (ASA) was extensively studied. Many of these principles were later adopted in the computational ASA (CASA) [12] approach.

It's important to point out that the majority of the works at that time exploit only the information in the current audio stream, which is very different from today's machine learning (ML)-based SOTA techniques that also take advantage of large training corpora collected or simulated over the time. Furthermore, most of the work at the time concentrated on monaural speech processing. This is because microphone arrays were considered expensive and were rarely used in practical systems, except in meeting scenarios. The single-microphone setup was believed to be more important and relevant. Both constraints limited the performance of the then-SOTA systems.

SLU and dialog systems

Interacting with machines in natural language has been of continued interest to mankind since the early days of computing. Around 1998, the popular architecture for dialog systems included language understanding, a dialog manager, and natural language generation. Language understanding aims to interpret user utterances into a semantic representation that can be converted to back-end knowledge and task-completion resources. Then, the dialog manager may formulate a query to the back end and predict the next system action based on the results of the query and the dialog context. Finally, natural

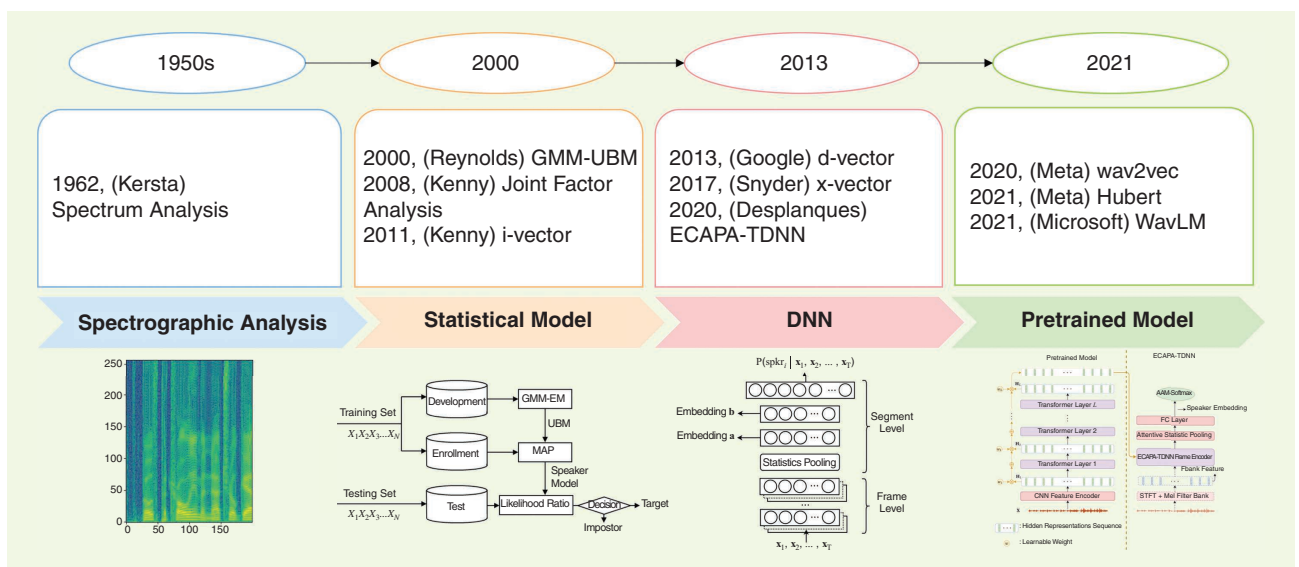


FIGURE 1. The progress of SPR systems. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation time-delay NN; STFT: short-time Fourier transform; Fbank: filter-bank; AAM: additive angular margin; FC: fully connected (layer); GMM-UBM: Gaussian mixture model MAP-adapted from a universal background model; DNN: deep NN; CNN: convolutional NN.

language generation produces the natural language utterance that conveys the system action.

For language understanding, approaches that rely on data and ML methods became feasible with the availability of annotated datasets, such as the air-travel-related queries of the DARPA airline travel information systems project. Traditionally, language understanding consists of a triage of tasks: intent/domain classification and slot tagging [identifying contiguous spans of words in an utterance that correspond to certain parameters (i.e., slots) of a user query], which have been treated as sequence classification and sequence-tagging problems, respectively [13]. The most popular technique for slot tagging at that time was conditional random field. For language-understanding tasks, annotated data enabled the combination or replacement of the earlier symbolic approaches with Bayesian classifiers and HMMs [14].

For dialog management, the common approaches around 1998 relied on dialog flows [15] that were designed by engineers to represent the interactions between the machine and humans. Approaches that rely on machine learning-based methods for learning dialog policies, such as [16], proposed reinforcement learning for predicting the next system action, were just starting to appear.

For language generation in dialog systems, the majority of the work was also template or grammar based. ML-based methods, which separated sentence planning and realization for language generation that aimed to reduce the high cost of handcrafting knowledge-based generation systems, started appearing.

Main driving forces over the last decades

SLP fields underwent a slow development period of roughly a decade and then went on to a fast track after 2010. Since 2010, we have seen rapid progress with various new modeling techniques and significantly improved performance. This progress is being driven by being able to relax previously existing modeling constraints through a combination of deep learning, big data, and high-performance computing.

Time for a paradigm change

Big data

The Internet and various digital applications significantly increased the number of data available to improve SLP systems. It was estimated that 2.5 quintillion bytes of data would be created every day in 2022. For example, in the 1998 time frame, typical large ASR systems would be trained on a few hundred hours of speech and 300 million words of text. Today, it is not uncommon to see systems trained on 100,000 h of speech, with some sites [17] using more than a million hours of speech. Although more data alone may not guarantee performance improvements, when combined with model-size increases, significant performance improvements result.

High bandwidth

Back in 1998, communication bandwidths were still at modem speeds (56 kb/s!), whereas today, bandwidth is on the order of

200 mb/s, even for a relatively low-end connection. This huge increase in bandwidth enables almost instantaneous uploading and downloading of speech signals and models, making it practical to utilize extremely large and accurate models in the cloud, resulting in dramatically increased SLP performance.

Affordable, high-performance computing

In a related development, the amount of computing power now available has also dramatically increased. Clock speed alone has increased by more than a factor of 100 over the last 25 years. The ability to pack multiple computing cores in a single processor and/or coprocessor has added even more computing capabilities. This enabled efficient parallelization of the large number of matrix operations required in deep learning. This allows for some very powerful SLP tasks to be done locally, providing enhanced latency, and also permits even bigger models to be employed in the cloud to achieve even greater performance gains and yet still run in real time.

Open source tools

Another driver of performance improvements has been the community's emphasis on research reproducibility and providing open source implementations of newly developed techniques, e.g., Kaldi (<https://github.com/kaldi-asr/kaldi>). Investments in deep learning platforms like TensorFlow and PyTorch have additionally sped up the rate of development of new speech processing toolkits, which powered a new wave of fundamental research, e.g., ESPnet (<https://github.com/espnet/espnet>), SpeechBrain (<https://github.com/speechbrain/speechbrain>), and Fairseq (<https://github.com/facebookresearch/fairseq>).

Deep learning and big models

Large-scale data sets and powerful computing infrastructure have enabled the adoption of deep learning techniques. Relative to the previous generation of technology, they have higher modeling capacity and can thus better leverage big data and offer significantly improved generalization abilities.

Neural architectures

Before the 2000s, single hidden layer, feedforward NNs were used in speech systems, either as a replacement for GMMs or as feature extractors [18], [19]. Later in the 2010s, deep feedforward NNs with many layers of latent representations began to replace latent variable models, e.g., GMMs, partially observable Markov decision processes, and i-vectors in speech systems, offering significant improvements across multiple benchmarks.

Unlike GMMs, NNs do not make any assumptions about the input data distribution. They are able to use far more data to constrain each parameter because the output for each training case is sensitive to a large fraction of the weights. Recurrent NNs (RNNs), with and without explicit memory, e.g., simple RNNs and long short-term memories, can capture long-range dependencies between inputs and bring finer-grain integration of temporal information for speech and language representation

[20]. In RNNs, connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes. However, the sequential nature of recurrent networks made it harder to leverage the all-parallel world of GPUs, which revolutionized the deep learning community with massive computational increases. Transformer layers [21] much better leverage GPUs by utilizing positionwise feedforward layers (i.e., the same feedforward layers are used for each positioned item in the sequence) and replacing the recurrence operation with a multihead self-attention sublayer (which runs through an attention mechanism several times in parallel) that captures input dependencies using a constant number of operations.

End-to-end modeling and optimization

Typical speech recognition, generation, translation, and dialog systems combine many logical components, e.g., feature representation, AM, LM, speaker, pronunciation, translation models, waveform generation, and hypothesis search, to name just a few. Modeling each logical component explicitly in classical systems ensured tight performance control and enabled easier integration of human knowledge. The first wave of research on neural models for spoken language systems represented each module with a neural counterpart, offering solid gains while maintaining the modularity of existing systems. Advances in numerical methods for optimizing NNs (e.g., layer normalization and residual connection) enabled neural models to combine multiple logical functions. Recent approaches trained single end-to-end models (instead of optimizing each component separately) to represent entire systems, e.g., ASR [20], [22], dialog systems [23], and TTS [24].

Big, pretrained models

One major bottleneck for deep neural models is their reliance on large volumes of labeled data, which is aggravated in end-to-end models that rely exclusively on data to skip low-level domain knowledge as feature engineering is accomplished automatically in the network instead of from human design. The help came from semisupervised (which utilizes both labeled and unlabeled data) and self-supervised training (which obtains supervisory signals from the data itself) approaches, which, by leveraging massive, unlabeled speech data, have reached unprecedented performance levels recently. Pseudolabeling, also known as *student–teacher distillation*, trains a student model on a few hours of labeled data to track outputs generated by a teacher model [25]. Self-supervised approaches [26] utilize pretext tasks to pretrain models generatively, contrastively, or predictively. These representation learning models impacted a wide range of downstream spoken language tasks, e.g., ASR, speaker diarization, SLU, and spoken question and answer.

Generative modeling

Parallel to the efforts in learning representations using unlabeled speech and audio data, there was an active line of research for modeling data distributions and learning high-quality generative models. Autoregressive (AR) generative models predict future values based on past values, both

given and predicted, in previous steps. They factorize a high-dimensional data distribution into a product of AR conditional distributions. Thanks to reparameterization, a way that solves undifferentiable expectations by rewriting them so that the distribution with respect to which we take the gradient is independent of model parameters, the encoder, which converts the input into a latent representation, and the decoder, which reconstructs the input from the latent representation, variational autoencoders (autoencoders whose input is encoded as a distribution over the latent space instead of a single point) can be trained jointly to reconstruct input data from samples of learned latent distributions. Generative adversarial networks (GANs) mimic the input data distribution through an adversarial game between a generator and a discriminator, aiming to discover realistic inputs from fake generated ones [27]. Utilizing these generative approaches has led to significant progress in generative models of speech that are controllable and of realistic quality.

Major technical breakthroughs in each subfield

Although deep learning has reshaped the whole field, different subfields have different problems to solve. In this section, we introduce the major technical breakthroughs in each major subfield in the past decades and describe the current SOTA.

Speech coding

As higher bandwidths became available for speech communication, first through Voice over IP applications such as voice and video conferencing over the Internet, and later in mobile communication such as Voice over LTE, bit rate scalable codecs were introduced. These codecs could operate at rates and bandwidths ranging from 5 to 6 kbps for narrow band and up to hundreds of kilobits per second for full-band speech and general stereo audio. Examples are Opus (<https://opus-codec.org/>), unified speech and audio coding [28], and enhanced voice services [29]. To achieve the ability to operate over this wide range of bit rates and signal bandwidths, these codecs combine linear predictive time-domain methods from low-rate speech coding with transform coding (such as the modified discrete cosine transform) common in high-quality general audio coding. Transform coding compresses audio data by representing the original signal with a small number of transform coefficients.

These are all still based on traditional digital signal processing techniques. As ML became successful in other speech processing areas, it also made its way into speech coding. WaveNet [30] showed that generative modeling can achieve impressive speech quality when conditioned on traditional speech analysis features such as spectral envelopes and pitch parameters. In [31], it was shown that traditional low-rate parametric speech codec features could drive a WaveNet neural synthesis and produce high-quality wideband speech at 2.4 kbps. Since then, other methods have been presented, driving down the complexity and bit rate further. Recent work [32] has produced high-quality speech at ~400 bps.

Most of today's neural speech codecs are excellent at reproducing clean speech, however, in a practical speech coding system, in addition to complexity constraints for running in

real time on devices, the system should also be able to handle speech in different background scenarios, e.g., in noisy and/or reverberant environments. This has been a challenge for neural speech codecs. In [33], an end-to-end system based on an autoencoder with GAN losses was proposed, and this method achieved good quality for both clean and noisy speech.

ASR

Initial advances in deep learning-based speech recognition were based on extensions to the existing GMM-HMM framework [34]. The basic HMM nature was not touched; deep learning was only applied to output distributions in the HMM framework. Over time, there were increasingly more attempts to replace the HMM framework with one based solely on deep learning.

A more speech-focused methodology called *connectionist temporal classification (CTC)* [35] combined HMM concepts with sequence-to-sequence mapping. It took almost 10 years to demonstrate competitive performance over hybrid models, with the realization that phone-based models worked better for CTC than state-based ones. CTC also produced a competitive performance with grapheme-based units, eliminating the need for costly pronunciation dictionaries (at least for systems with adequate numbers of training data).

Benefiting from the monotonic relationship between ASR inputs and outputs, the RNN transducer [20] took the modeling process further by augmenting the AM with a prediction network, which replaced the need for an LM and was trained jointly within the whole “end-to-end” ASR model.

Encoder–decoder models with attention [21], a processing mechanism that allows a network unit at a layer to pay more attention (with a higher weight) to specific units at other layers, were then successfully adapted from the translation community, but such networks’ freedom to reorder outputs would sometimes introduce new types of speech recognition errors.

Another significant advance in speech recognition occurred when it was realized that self-supervised learning concepts, as embodied in bidirectional encoder representations from transformers, could be adapted to improve speech recognition performance. To achieve that, the concept of masking discrete elements in a text stream, as is done in Bidirectional Encoder Representation from Transformers (BERT), needed to be extended to speech, which is a continuous signal. More generally, self-supervised methods were further extended to the pretraining of speech models [26] so that the more accessible, unlabeled speech data may be exploited to improve speech processing performance. Again, the main challenge here was extending models originally developed for discrete units (text) to continuous units (speech) without obvious reconstruction targets, as there are in a text stream.

TTS

Unit-selection TTS was popular in R&D in the early 2000s. There were many commercial unit-selection TTS systems and

several open source software toolkits. In generative TTS, statistical parametric speech synthesis [36] with high-quality vocoders gained popularity in the late 2000s. The Blizzard Challenge (https://www.synsig.org/index.php/Blizzard_Challenge), an annual event that evaluates TTS systems with a common training speech corpus and a set of test sentences, started in 2005 and helps researchers and developers compare different technologies on the same ground.

Deep learning was first introduced to replace the HMM-based AM in generative TTS [37]. In 2016, WaveNet [30], an AR generative model for raw audio, demonstrated that it could integrate an AM and a vocoder into a single generative model and synthesize more naturally sounding speech than conventional unit-selection and generative TTS systems. In parallel, the AR encoder–decoder models with attention were successfully adapted as an AM of generative TTS [38] like other sequence-to-sequence mapping problems. A combination of the encoder–decoder model with the WaveNet-based vocoder model achieved near-human-level synthetic speech [39]. Recently, non-AR generative models demonstrated that they could achieve the same or better performance than these AR generative models, both in AMs and vocoders [1]. Finally, integrating these two components into a single model to make the entire system fully end to end is being actively investigated [24].

Some SOTA, NN-based, TTS systems have demonstrated human parity in the reading-speech (in contrast to conversational speech that features wide prosody variations) synthesis domain. Current research in TTS targets harder speech generation tasks, such as synthesizing texts in low-resource languages, handling code mixing (the embedding of linguistic units such as words and morphemes of one language into an utterance of another language), code switching (alternating between two or more languages) within a sentence, synthesizing long-form texts, realizing expressiveness, and synthesizing nonverbal vocalizations such as laughter. Humans are still significantly better than TTS with these tasks. Developments in data collection, analysis, and modeling can help us tackle these hard tasks.

SPR, identification, and diarization

In SPR, the beginning of 2000s was dominated by Gaussian mixture speaker models MAP-adapted from a universal background model (GMMs-UBM) [40]. Two important variations of using adapted GMMs existed: 1) direct evaluation of the likelihood of utterance, where the verification score was the log-likelihood ratio between GMM-UBM and speaker-adapted GMM, and 2) extracting adapted GMMs’ parameters as “speaker supervectors” and using them as the input to another classifier [e.g., a support vector machine (SVM)].

Many techniques, such as feature mapping and nuisance attribute projection, were developed to compensate for channel variability so that speaker variability could be better identified. In [41], joint factor analysis (JFA) was introduced as an improvement to the previous GMM-UBM/MAP approach,

A more speech-focused methodology called connectionist temporal classification combined HMM concepts with sequence-to-sequence mapping.

where large GMM models could be robustly and independently adapted to the speaker and/or channel of an utterance. Similar to the eigenvoices adaptation used in ASR, in which each speaker is represented as a linear combination of latent basis vectors named *eigenvoices*, only low-dimensional speaker and channel latent vectors needed to be estimated from an input utterance. The subsequent i-vector approach directly used the latent vectors as low-dimensional, fixed-length embeddings of speech utterances [42]: i-vectors defined only one “total-variability” space, and supervectors of GMM means were projected into such a space by a total-variability matrix trained on a huge number of unlabeled speaker data. I-vectors, however, included both wanted speaker and unwanted channel information, so scoring had to be implemented by probabilistic linear discriminant analysis (PLDA), rather than by simple distance metrics. I-vectors dominated the field for more than a decade, and they became popular elsewhere, from LR to the adaptation of ASR systems (even those based on deep learning).

SPR was actually one of the last ML fields where GMMs surrendered to NNs. Efforts have been made for more than a decade, and researchers have registered partial victories (such as NN-based features and NN alignments, instead of using Gaussian components), but the true switch to NNs came after Snyder et al. [43] trained a time-delay NN on a large pool of speakers with a speaker identification criterion. The NN has several blocks: 1) extracting frame-by-frame hidden representations; 2) pooling over time, resulting in a fixed-length representation of an utterance; 3) adding a few more NN layers to produce the embedding (x-vector); and 4) during training, the “x-vector” is connected to a linear classification layer, which was discarded once the “x-vector” extractor was trained. Since the introduction of x-vectors, the SPR standard architecture has stabilized with the chain feature-extraction—embedding extraction—back end. Current work in SPR is compatible with the other ML fields and includes research in data augmentation, novel network architectures (often taken from the computer vision community, such as ResNet34), training criteria, end-to-end systems (including trainable signal processing blocks), and the use of pretrained models.

In diarization, the Bayesian information criterion (BIC) [44] has long been used for both segmentation and clustering. Diarization has closely followed the developments in SPR: i-vectors (or x-vectors) were used to represent the speech segments, and PLDA was used to evaluate the similarity for segment clustering. Also, the BIC-based segmentation was replaced by simpler uniform segmentation, where i- and x-vectors are extracted every 0.25 s from a window of approximately 1.5 s. Hierarchical agglomerative clustering, spectral clustering, or clustering based on Bayesian HMMs were typically used to cluster the segments. Variational Bayes (VB) diarization (Bayesian HMM with eigenvoice priors [45]) was unique as it did not perform separate segmentation and clustering steps. VB techniques worked excellently with deep NN (DNN)-generated x-vectors representing segments of fixed length. Current work in diarization also targets end-to-end architectures, and promising results have been obtained with target-speaker voice activity detection.

LID

Significant improvements were made to the acoustic approach of LID in the early 2000s by reusing discriminative training that had previously been tested in ASR. As phone recognition was the first speech field where NNs achieved significant success, it is not surprising that the phonotactic approach of LID benefited from the development of reliable phone recognizers. LID then evolved alongside SPR because the same groups and people typically worked on both techniques, and some of this evolution is covered by [46]. LID based on JFA and i-vectors virtually eliminated the need for phonotactic approaches.

Although some attempts to use DNNs for LID still found it advantageous to combine with GMMs and use NNs as feature extractors, others have shown the superiority of NNs, leading to neural approaches dominating the LID field earlier than SPR. X-vectors have also proven their modeling power for LID [47], with simple, discriminative Gaussian classifiers used as the back end. The interest in LID also initiated several data collection efforts, from the extraction of telephone calls from broadcasts done by Brno University of Technology and the Linguistic Data Consortium around 2009, to the recent VoxLingua107 data collection. As for SPR, LID currently witnesses developments in data augmentation, new NN architectures, and end-to-end systems.

Speech enhancement and separation

Over the past two decades, we have observed significant progress in speech enhancement and separation. Most of the developments are summarized in [48], [49], [50], [51], and [52].

In 2001, nonnegative matrix factorization (NMF) [52], an unsupervised data-driven technique, was introduced under the assumption that the audio spectrogram has a low rank structure that can be represented with a small number of nonnegative bases. Under certain conditions, the decomposition in NMF is unique, without making orthogonality or independence assumptions. The main difference between NMF and previous signal model-based approaches (e.g., the Wiener filter) is that NMF uses clean speech and noise streams to learn the basis, and then applies these bases during the testing phase.

Around the same time, CASA was proposed [12]. In CASA, certain segregation rules based on perceptual grouping cues (e.g., pitch and harmonics that can be used to distinguish different speakers) are designed to operate on low-level features such as a spectrogram to estimate a time-frequency (T-F) mask for each signal component belonging to different speakers. This mask is then used to reconstruct the signal by multiplying it with the input. Although CASA has many limitations [50], the idea of estimating T-F masks, when combined with data-driven approaches, has reshaped the direction of speech enhancement.

Deep learning has also led to a paradigm shift in speech enhancement and separation. The key idea is to convert the original problem into a supervised learning problem [49], [50]. As the target clean speech is seldom available in real-world recordings, the training sets are usually synthesized by mixing various clean speech, noise, and reverberation conditions. The task then becomes extracting the clean speech from the synthesized mixture. As the mixing sources and parameters are known during

the training phase, various training objectives (mostly T-F masks and signal matching based) can thus be directly defined if only one speaker needs to be extracted from the mixture. However, if two or more speakers need to be separated and extracted from the mixture and there is no information (e.g., speaker embedding, face, or location) during the segregation process to identify the order of the set of extracted speech streams, some technique is needed to solve the permutation-ambiguity issue [50]. The two most effective approaches for solving this problem are deep cluster [53] and permutation-invariant training (PIT) [54]. Unfortunately, the synthesized mixture, although it can be close, is different from the real recordings. To exploit the real recordings that do not come with the separation or enhancement targets, mixture-invariant training [55] was proposed and achieved significant success when combined with PIT.

Another observation in the past two decades is the improved exploitation of additional information. For example, the research on multichannel processing [48] and multimodal processing [51] has significantly increased. Multichannel processing can utilize spatial information to improve the performance of speech enhancement and separation. Beamforming based solely on signal processing was the dominant multichannel technique 10 years ago. Today, deep learning has been exploited to estimate sound statistics, learn a dynamic beamformer, and introduce additional target clues such as speaker embedding and multimodal information.

SLU and dialog systems

The past two decades have been flourishing for dialog systems. Due to the advancements in speech and language technology, several commercial applications, such as customer service applications, virtual personal assistants, and social bots, have been launched, resulting in a huge number of interactions. These have resulted in even more research interest in dialog systems as they have enabled us to identify remaining challenges and new conversational application domains and scenarios.

For language understanding, the methods relying on SVMs and CRFs were followed by the use of DNNs [56] and large, pretrained LMs fine-tuned for language understanding [57]. Similar to language understanding, pretrained LMs have proven to be useful for other dialog tasks as well; for example, dialog-state tracking and response generation. Inspired by these works, zero- (with no training sample) and few-shot (with only several training samples) methods that rely on fine-tuning question answering [58], prompt tuning, and instruction tuning [59] were shown to be effective.

In parallel, end-to-end methods based on deep learning [23], for both task-oriented and open-domain dialog systems, became popular. Most recently, ChatGPT, a general-purpose, open-domain dialog system based on the generative predictive transformer [60], has shown great potential and become prevalent.

Conclusion

In this article, we reviewed major advances made in the SLP fields in the past 25 years. The availability of more data, higher

computation power, and advancements in deep learning techniques have accounted for the majority of the progress made in SLP in the last decade. In this section, we summarize the state of the field today, with comments on future developments.

Comments on the field today

Figure 2 compares the ICASSP SLP paper theme category percentages (2023 versus 1998). We observe that the percentage of categories such as ASR, speech synthesis, SPR, language modeling, speech enhancement, and speech analysis has drastically increased over the last 25 years. Language understanding, emotion recognition, voice conversion, multilingual ASR, speaker diarization, speech corpora and resources, ML for language processing, and self-supervised learning emerge as significant theme categories. In contrast, ASR robustness (now achieved with a large number of realistic training data), features for ASR (feature engineering is now part of data-driven model training), NNs (speech modeling with NNs has become a universal tool), and speech coding no longer take up a noticeable percentage as theme categories.

Figure 3 shows the evolution of ICASSP SLP paper submission statistics for the past 20 years. We observe that the number of paper submissions has nearly tripled. Especially, we see a roughly 20%/year-over-year increase for the last five years.

The developments in SLP have enabled many scenarios and significantly improved our daily life. For example, ASR techniques, given their significantly improved accuracy, are now widely used in smart assistants such as Siri, Alexa, and Google Now; in-car entertainment systems, voice search systems, and medical transcription systems. ASR techniques have also enabled many other scenarios, such as speech-to-speech translation. Due to the high quality of synthesized speech, TTS techniques have significantly improved the multimodal and multimedia context generation process. They are widely used in audiobooks, digital humans, and dialog systems.

Perspectives on future developments

We have observed the convergence of techniques in most of the subfields of SLP. Only decades ago, these subfields were based on very different theories and models; today, most of them are based on the same set of deep learning techniques. When a new effective model or algorithm is developed in one of these subfields, it is quickly applied to other subfields and brings progress to them. We welcome this trend and believe it will continue because, although many problems seem to be different on the surface, they are identical at a higher and more abstract level. At the same time, we believe problems in different subfields come with different assumptions and have different structures. These assumptions and structures should be taken into consideration when designing models to advance the SOTA.

Given this convergence, one of the main impacts we expect to see in the coming years is an increasing number of systems that may have different preprocessing and postprocessing modules for different modalities, but share a common central architecture. This will facilitate cross-modal learning and data

sharing, something humans do easily but with which current systems still struggle.

For language understanding and dialog-response generation, although these new methods (e.g., ChatGPT) have resulted in significant improvements, several challenges remain, such as maintaining the factual accuracy of the responses,

modeling longer context that is important for future interactions, and common-sense reasoning.

Another area that still requires significant technology advancement deals with the general area of catastrophic forgetting. The SOTA today in speech and other modalities involves fine-tuning a large, pretrained network; this results in biasing

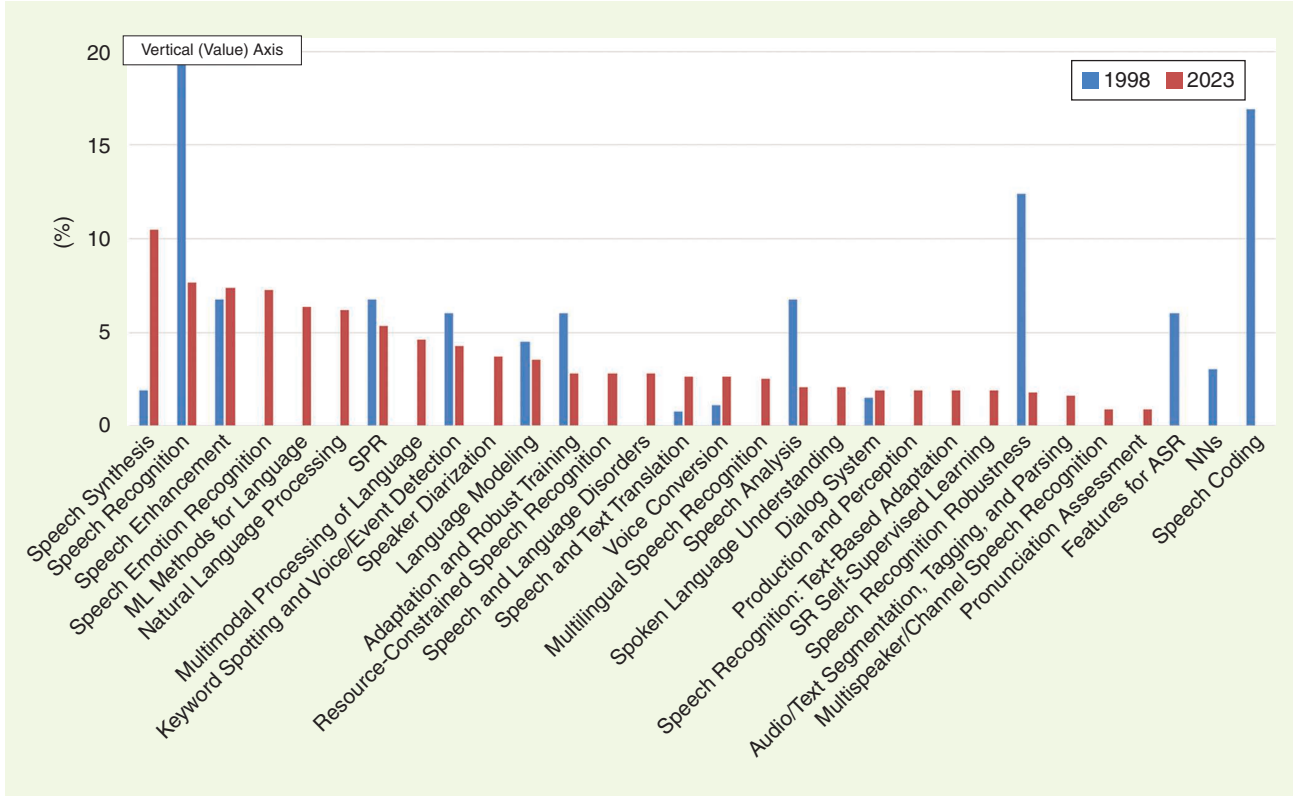


FIGURE 2. Evolution of the percentage of paper category of ICASSP SLP papers (1998 versus 2023), ordered after 2023 category percentage.

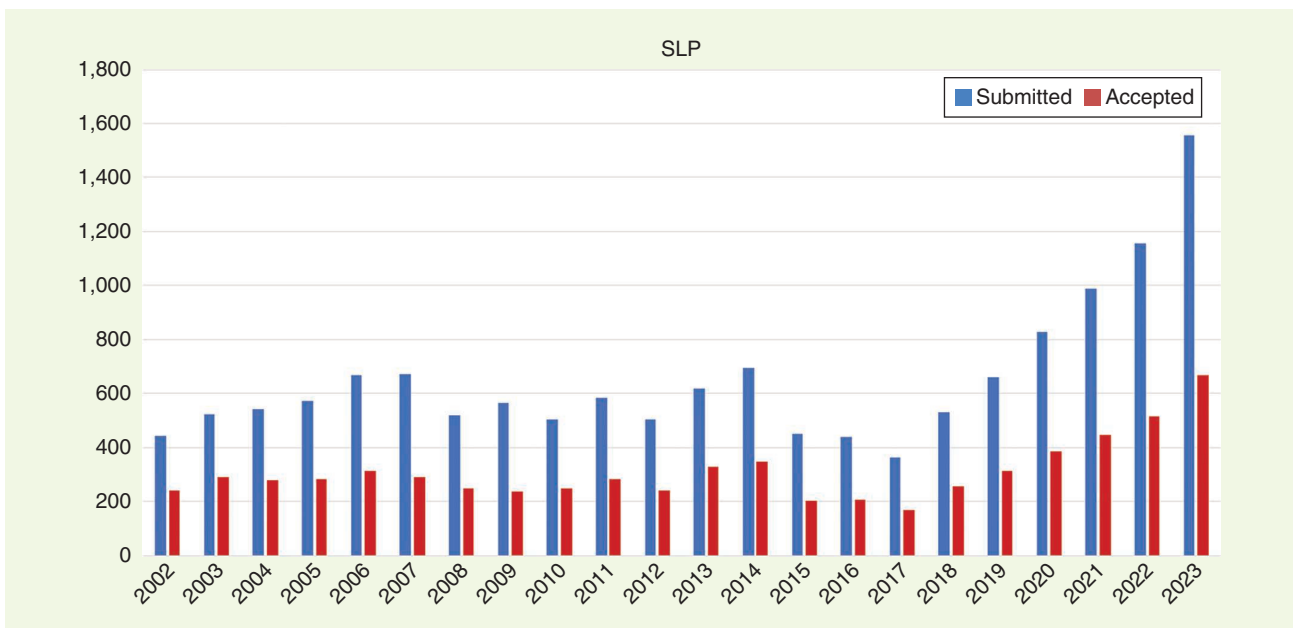


FIGURE 3. Evolution of ICASSP SLP paper submission statistics (from 2002 to 2023).

the network to the new data with a resultant loss in robustness. Again, although there have been some attempts to deal with this problem, we have a long way to go.

We'd also like to point out that the majority of the progress has been made in support of data-driven techniques. This, however, does not mean that theoretical models are useless or less meaningful. In fact, we think theories on deep learning should be established to explain the models and results we have thus far, and new theories should be developed to guide further development of the fields in which they apply. It is also beneficial to combine theoretical models with data-driven techniques, e.g., in the speech enhancement and separation fields.

As the performance of the systems continues to improve, it is very important to maintain ethics. For example, as synthesized speech is no longer distinguishable from human speech under some conditions, we need to make sure that the advanced TTS technique is not used to cheat people or for illegal financial gain. The research society should have clear guidance on how to evaluate the benefits and side effects of new research and techniques. Both technical and legal mechanisms are also needed to prevent evil people from getting access to powerful techniques, and to identify AI-generated content.

Acknowledgment

Dong Yu is the corresponding author.

Authors

Dong Yu (dongyu@ieee.org) received his Ph.D degree in computer science from the University of Idaho. He is a distinguished scientist and vice general manager at Tencent AI Lab, Bothell, WA 98011 USA. Prior to joining Tencent in 2017, he was a principal researcher at Microsoft Research, Redmond, WA. He has two monographs and more than 300 papers to his credit. His work has been widely cited and recognized by the IEEE Signal Processing Society Best Paper Award in 2013, 2016, 2020, and 2022, as well as by the 2021 NAACL Best Long Paper Award. He was elected chair of the IEEE Speech and Language Processing Technical Committee from 2021 to 2022. His research focuses on speech and natural language processing. He is a Fellow of IEEE and a fellow of Association for Computing Machinery and the International Speech Communication Association.

Yifan Gong (yifan.gong@ieee.org). received his Ph.D. degree in computer science from the Department of Mathematics and Computer Science, University of Nancy I, France. He leads a speech modeling team developing machine learning and speech technologies across scenarios/tasks at Microsoft, Redmond, WA USA. Prior to joining Microsoft in 2004, he worked as a senior research scientist at the National Center of Scientific Research, France, and then as a senior member of the technical staff at Texas Instruments. He has authored and coauthored more than 300 publications in books, journals, and conferences, and has more than 70 granted patents. He serves on the Senior Editorial Board of *IEEE Signal Processing Magazine* and is the elected chair of the IEEE Speech and Language Processing Technical Committee

(2023–2024). His research focus is on speech processing. He is a Fellow of IEEE.

Michael Alan Picheny (map22@nyu.edu) received his Sc.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology. He has worked in the speech recognition area since 1981, joining the IBM Thomas J. Watson Research Center after finishing his doctorate degree. After retiring from IBM in 2019, he joined NYU-Courant Computer Science and the Center for Data Science, New York, USA, as a part-time research professor. He has published numerous papers in both journals and conferences on almost all aspects of speech recognition. He was chair of the IEEE Speech Technical Committee from 2002 to 2004 and a member of the International Speech Communication Association (ISCA) Board from 2005 to 2014. Her research focus is on speech recognition. He is a Fellow of IEEE and a fellow of the ISCA.

Bhuvana Ramabhadran (bhuv@google.com) received her Ph.D. degree in electrical engineering from the University of Houston. She leads a team of researchers focusing on semisupervised learning for speech recognition and multilingual speech recognition at Google, New York, NY 10011 USA. Prior to joining Google, she was a distinguished research staff member and manager at IBM Research AI, IBM Thomas J. Watson Research Center, New York, where she led a team of researchers in the Speech Technologies Group and coordinated activities across IBM's worldwide laboratories in the areas of speech recognition, synthesis, and spoken-term detection. She was elected chair of the IEEE Speech and Language Processing Technical Committee from 2015 to 2016. Her research interests include speech recognition and synthesis algorithms, statistical modeling, signal processing, and machine learning. She is a Fellow of IEEE and a fellow of the International Speech Communication Association.

Dilek Hakkani-Tür (dilek@ieee.org) received her Ph.D. degree in computer engineering from Bilkent University. She is a senior principal scientist focusing on enabling natural dialogs with machines at Amazon Alexa AI, Sunnyvale, CA USA, Prior to joining Amazon, she was a researcher at Google, Microsoft Research, the International Computer Science Institute at the University of California, Berkeley, and AT&T Labs-Research. She received best paper awards for publications she coauthored on conversational systems from the IEEE Signal Processing Society, International Speech Communication Association (ISCA), and European Association for Signal Processing. Recently, she served as editor-in-chief of *IEEE Transactions on Audio, Speech, and Language Processing*. Her research interests include conversational artificial intelligence, natural language and speech processing, spoken dialog systems, and machine learning for language processing. She is a Fellow of IEEE and a fellow of the ISCA.

Rohit Prasad (roprasad@amazon.com) received his M.S. degree in electrical engineering from the Illinois Institute of Technology. He is a senior vice president at Amazon, Boston, USA, where he is head scientist for Amazon Alexa. He leads R&D in artificial intelligence technologies for enriching the

daily lives of everyone, everywhere. Prior to Amazon, he was the deputy manager and senior director of the Speech, Language and Multimedia Business Unit at Raytheon BBN Technologies. He is a named author on more than 100 scientific articles and holds several patents. He was listed at number nine in Fast Company's "100 Most Creative People in Business" in 2017 for leading the "voice-controlled revolution." In 2021, he was listed as one of the 50 most influential people in technology as part of "Future Tech Awards." His research interests include speech processing and dialog system. He is a Senior Member of IEEE.

Heiga Zen (heigazen@google.com) received his Ph.D. degree in computer science from the Nagoya Institute of Technology in 2006. He is a research scientist with the Google Brain team, Tokyo 150-0002, Japan. After receiving his Ph.D. degree, he joined Toshiba Cambridge Research Laboratory, U.K., in 2008. He was with the Text-to-Speech team at Google, London, between 2011 and 2018, then moved to the Brain team in Tokyo as one of its founding members. He is one of the early explorers in generative model-based speech synthesis, one of the original authors of the hidden Markov model-based speech synthesis system HMM/DNN-based Speech Synthesis System (HTS), and its first maintainer. He served as a member of the IEEE Speech and Language Processing Technical Committee between 2012 and 2014. He is a fellow of the International Speech Communication Association. He is a Senior Member of IEEE.

Jan Skoglund (jks@google.com) received his Ph.D. degree in information theory from the School of Electrical and Computer Engineering of Chalmers University of Technology, Sweden, in 1998. He leads a team that develops speech and audio signal processing components for capture, real-time communication, storage, and rendering (deployed in products such as Meet and Chromebooks) at Google, San Francisco CA USA. After receiving his Ph.D. degree, he worked on low bit rate speech coding at AT&T Labs-Research, Florham Park, NJ. He was with Global IP Solutions (GIPS), San Francisco, from 2000 to 2011, working on speech and audio processing tailored for packet-switched networks. GIPS' audio and video technology is found in many deployments by, e.g., IBM, Google, Yahoo, WebEx, Skype, and Samsung, and was open sourced as WebRTC after a 2011 acquisition by Google. His research interests include speech and audio signal processing. He is a Senior Member of IEEE.

Jan "Honza" Černocký (cernocky@fit.vutbr.cz) received his Ph.D. degree in electronics from the University Paris XI. He is a full professor and head of the Department of Computer Graphics and Multimedia, Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic. He also serves as managing director of the BUT Speech@FIT research group. He is responsible for signal and speech processing courses at FIT BUT. In 2006, he cofounded Phonexia. He was general chair of INTERSPEECH 2021 in Brno. His research interests include artificial intelligence, signal processing, and speech data mining (speech, speaker, and language recognition). He is a Senior Member of IEEE.

Lukáš Burget (burget@fit.vutbr.cz) received his Ph.D. degree in information technology from Brno University of Technology. He is an associate professor with the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Brno, Czech Republic, and research director of the BUT Speech@FIT group. He was a visiting researcher with OGI Portland and SRI International. He was on numerous European Union- and U.S.-funded projects and currently leads the NEUREM3 project, which is supported by the Czech Science Foundation. In 2006, he cofounded Phonexia. His research interests are in the field of speech processing, concentrating on acoustic modeling for speech, speaker, and language recognition. He is a Member of IEEE.

Abdelrahman Mohamed (manosami@gmail.com) received his Ph.D. in computer science from the University of Toronto. He is a research scientist with the Facebook Artificial Intelligence Research group at Meta, Seattle, WA, USA. Prior to joining Meta, he was a principal scientist/manager at Amazon Alexa AI and a researcher at Microsoft Research. He was a part of the team that started the deep learning revolution in spoken language processing in 2009. He has more than 70 research journal and conference publications with more than 35,000 citations. He is the recipient of the 2016 IEEE Signal Processing Society Best Journal Paper Award. He currently serves as a member of the IEEE Speech and Language Processing Technical Committee. His research work spans speech recognition; representation learning using weakly, semi-, and self-supervised methods; language understanding; and modular deep learning. He is a Member of IEEE.

References

- [1] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2020, *arXiv:2006.04558*.
- [2] L. Supplee, R. Cohn, J. Collura, and A. McCree, "MELP: The new federal standard at 2400 bps," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, vol. 2, pp. 1591–1594, doi: 10.1109/ICASSP.1997.596257.
- [3] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1985, pp. 937–940, doi: 10.1109/ICASSP.1985.1168147.
- [4] F. Jelinek, "Some of my best friends are linguists," *Lang. Resour. Eval.*, vol. 39, no. 1, pp. 25–34, Feb. 2005, doi: 10.1007/s10579-005-2693-4.
- [5] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR v -talk speech synthesis system," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 1992, pp. 483–486.
- [6] R. Donovan and P. C. Woodland, "Automatic speech synthesiser parameter estimation using HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1995, pp. 640–643, doi: 10.1109/ICASSP.1995.479679.
- [7] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1995, pp. 660–663, doi: 10.1109/ICASSP.1995.479684.
- [8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, pp. 1611–1614, doi: 10.1109/ICASSP.1997.598807.
- [9] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997, doi: 10.1109/5.628714.
- [10] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, p. 31, Jan. 1996, doi: 10.1109/TSA.1996.481450.
- [11] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.
- [12] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004, doi: 10.1109/TNN.2004.832812.

- [13] G. Tur and R. De Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information From Speech*. Hoboken, NJ, USA: Wiley, 2011.
- [14] W. Minker, S. Bennacef, and J.-L. Gauvain, "A stochastic case frame approach for natural language understanding," in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, vol. 2, pp. 1013–1016, doi: 10.1109/ICSLP.1996.607775.
- [15] P. C. Constantinides, S. Hansma, C. Tchou, and A. I. Rudnick, "A schema based approach to dialog control," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP)*, 1998, Paper 0637, doi: 10.21437/icslp.1998-68.
- [16] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 1, pp. 11–23, Jan. 2000, doi: 10.1109/89.817450.
- [17] S. H. K. Parthasarathi and N. Ström, "Lessons from building acoustic models with a million hours of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 6670–6674, doi: 10.1109/ICASSP.2019.8683690.
- [18] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer, 1993.
- [19] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000, vol. 3, pp. 1635–1638, doi: 10.1109/ICASSP.2000.862024.
- [20] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, vol. 30.
- [22] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.
- [23] R. T. Lowe, N. Pow, I. Serban, L. Charlin, C.-W. Liu, and J. Pineau, "Training end-to-end dialogue systems with the ubuntu dialogue corpus," *Dialogue Discourse*, vol. 8, no. 1, pp. 31–65, Jan. 2017, doi: 10.5087/dad.2017.102.
- [24] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5530–5540.
- [25] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014.
- [26] A. Mohamed et al., "Self-supervised speech representation learning: A review," 2022, *arXiv:2205.10643*.
- [27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, vol. 27, pp. 2672–2680.
- [28] M. Neuendorf et al., "The ISO/MPEG unified speech and audio coding standard – Consistent high quality for all content types and at all bit rates," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 956–977, Dec. 2013.
- [29] M. Dietz et al., "Overview of the EVS codec architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 5698–5702, doi: 10.1109/ICASSP.2015.7179063.
- [30] A. van den Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
- [31] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "Wavenet based low rate speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 676–680, doi: 10.1109/ICASSP.2018.8462529.
- [32] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," 2021, *arXiv:2104.00355*.
- [33] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 495–507, 2022, doi: 10.1109/TASLP.2021.3129994.
- [34] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.
- [35] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376, doi: 10.1145/1143844.1143891.
- [36] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009, doi: 10.1016/j.specom.2009.04.004.
- [37] Z.-H. Ling, S. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, pp. 35–52, May 2015, doi: 10.1109/MSP.2014.2359987.
- [38] J. Sotelo, S. Mehri, K. Kumar, J. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proc. ICLR Workshop*, 2017.
- [39] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.
- [40] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. II-53–II-56, doi: 10.1109/ICASSP.2003.1202292.
- [41] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008, doi: 10.1109/TASL.2008.925147.
- [42] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. Interspeech*, 2009, 1559–1562, doi: 10.21437/Interspeech.2009-385.
- [43] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.
- [44] S. Chen et al., "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, vol. 8, pp. 127–132.
- [45] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on Bayesian hmm with eigenvoice priors," in *Proc. Odyssey*, 2018, pp. 147–154.
- [46] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013, doi: 10.1109/JPROC.2012.2237151.
- [47] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Proc. Odyssey*, 2018, pp. 105–111.
- [48] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017, doi: 10.1109/TASLP.2016.2647702.
- [49] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018, doi: 10.1109/TASLP.2018.2842159.
- [50] Y.-M. Qian, C. Weng, X.-K. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 40–63, Jan. 2018, doi: 10.1631/FITEE.1700814.
- [51] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1368–1396, Mar. 2021, doi: 10.1109/TASLP.2021.3066303.
- [52] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, vol. 13, pp. 556–562.
- [53] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 31–35, doi: 10.1109/ICASSP.2016.7471631.
- [54] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 241–245, doi: 10.1109/ICASSP.2017.7952154.
- [55] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, vol. 33, pp. 3846–3857.
- [56] G. Mesnil et al., "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 3, pp. 530–539, Mar. 2015, doi: 10.1109/TASLP.2014.2383614.
- [57] Q. Chen, Z. Zhuo, and W. Wang, "BERT for joint intent classification and slot filling," 2019, *arXiv:1902.10909*.
- [58] M. Namazifar, A. Papangelis, G. Tur, and D. Hakkani-Tür, "Language model is all you need: Natural language understanding as question answering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 7803–7807, doi: 10.1109/ICASSP39728.2021.9413810.
- [59] P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. P. Bigham, "Improving zero and few-shot generalization in dialogue through instruction tuning," 2022, *arXiv:2205.12673*.
- [60] OpenAI, "GPT-4 technical report," 2023. [Online]. Available: <https://arxiv.org/abs/2303.08774>