From KITT to Amazon Alexa: A primer to automatic speech recognition and text-to-speech with applications in Marketing.

Francesc Busquet

Institute of Behavioral Science and Technology, Torstrasse 25, 9000 St. Gallen, Switzerland

E-mail: francesc.busquet@unisg.ch

ORCID: 0000-0002-2316-7722

Abstract

This paper explores the potential of automatic speech recognition and text-to-speech technologies in the field of marketing. From the fictional world of Knight Rider's KITT to the reality of Amazon Alexa, these technologies have come a long way and now offer companies new opportunities to create value on a large scale. Therefore, this paper provides a comprehensive overview of the evolution, challenges, and practical applications of automatic speech recognition and text-to-speech technologies in marketing, making it a valuable resource for marketers looking to further understand the operations of these technologies, and how to leverage them to enhance consumer experience and engagement. Finally, an appendix provides a compilation of relevant tools and resources to aid interested readers in furthering their understanding and implementation of these cutting-edge technologies.

Keywords: automatic speech recognition, text-to-speech, language, speech

From KITT to Amazon Alexa: A primer to automatic speech recognition and text-to-speech with applications in Marketing.

Speech is the most natural form of communication in human interaction, being the primary way by which people communicate their ideas and emotions, and essential for social interaction.

With the generalization of the first computational systems, more than fifty years ago, people started to envision computers that could easily hold a conversation. A clear example of that can be seen in the 1982 TV series Knight Rider, where an artificially intelligent car, named KITT, is capable of humanlike conversation.

Despite that, such examples were limited to fiction, as text was the only means of interacting with computers and it is still the prevailing means to interact with them. Nevertheless, the desire to achieve such a goal never vanished, and researchers focused on investigating how to give ears and mouths to computers. The results of these investigations brought continuous advances in automatic speech recognition (ASR), which allows the conversion of speech into text, and text-to-speech (TTS), which enables the opposite, i.e., the conversion of text to speech.

These technologies are still far from matching human performance. Nevertheless, they are already being successfully used in numerous applications. A recent example of such applications are voice assistants like Amazon Alexa or Google Assistant, which are software agents that can interpret human speech and respond via synthesized voices. By allowing users to control devices and access information hands-free, voice assistants bring the fictional concept of KITT a step closer to reality. With consumers readily adopting these technologies and already making purchases with them, it's clear that voice assistants are changing the way we interact with technology, and brands should leverage such technologies (Capgemini Digital Transformation Institute, 2018).

Thus, ASR and TTS provide a new paradigm to interact with computers and other devices. In addition, they open the window to interact with consumers in fully automated ways and to better understand them. In the consumer service domain, for example, ASR can be used to automatically transcribe consumer calls, which can then be analyzed to extract information that can be used to improve the service. Likewise, TTS can be used to generate automated responses to consumers, reducing the need for human operators.

For this reason, in this paper, we present a marketing-focused examination of the advancements, promises, and challenges of ASR and TTS technologies. By providing a fundamental understanding of the operations of these technologies, we aim to provide marketers with a deeper understanding of how to effectively leverage ASR and TTS to improve consumer experience and engagement. For readers looking to dive deeper, an appendix has been included that provides links to relevant tools and readings.

1. Automatic speech recognition

ASR, also known as speech-to-text, is the process of converting spoken language into text. Despite this task may seem trivial to humans, it is nonetheless a very challenging task for computers. This is because humans can effortlessly recognize and interpret speech, while computers must analyze speech waveforms to identify the units of sound that make up the utterance. This process is complicated by the fact that there can be a great deal of variation in the way that people produce speech sounds (Benzeghiba et al., 2007). For example, the same word can be pronounced differently by different people, or even by the same person at different times. This variability makes it difficult for computers to accurately identify the sounds that makeup speech.

1.1. The early stages of automatic speech recognition

The first automatic speech recognition system was developed in 1952 by a team of researchers at Bell Laboratories (Davis et al., 1952). This system was only able to recognize digits from zero to nine. Despite that, it was a breakthrough at the time and paved the way for further research in this area. During the following years, several other systems appeared mostly for isolated word recognition (e.g., Denes & Mathews, 1960; King & Tunis, 1966; Martin, Nelson, & Zadell, 1964; Vonusa, Jacobs, & Beek, 1970).

Most of these early systems were based on simplistic template-matching approaches, where each spoken utterance was compared to a set of stored templates. If the match was close enough, the system would output the corresponding word. These models proved to be too simplistic, working well for isolated word recognition but not being very effective for continuous speech recognition (O'Shaughnessy, 2008). As early as 1966, forward-thinking entities recognized the potential of ASR to enhance consumer experiences. One notable example is the Informedia Digital Video Library Project of Carnegie Mellon University, which planned the use of ASR to facilitate video searches (de Silva & McKnight, 1966).

1.2. Development of more sophisticated approaches using hidden Markov models

With the advent of more powerful computers in the 1970s, researchers began to explore more sophisticated approaches to ASR. One of the most promising approaches was based on Hidden Markov Models (HMMs) (L. Rabiner & Juang, 1986), which proved to be much more successful at recognizing continuous speech (Juang & Rabiner, 1991). Due to these improvements, some authors already discussed the feasibility of ASR in serving consumers (e.g., Flanagan, 1976; Holmgren, 1983; Wattenbarger et al., 1993).

A HMM is a statistical model that can be used to predict the probability of a sequence of hidden states, which are dependent on each other (Eddy, 2004). They are hidden in the sense that they are not directly observable, instead, we can observe some variables which are related to them and by using these observed variables, we can predict the sequence of hidden states.

The usage of HMMs in ASR is based on the idea that speech can be understood as a sequence of hidden states, i.e., the phonemes in the spoken utterance, which can be inferred from observed variables, i.e., the acoustic features of the speech signal (L. R. Rabiner, 1989; Waibel & Lee, 1990).

There are several different approaches to HMM-based speech recognition, but the basic architecture of an HMM-based speech recognition system consists of two main steps: (1) feature extraction and (2) decoding (Gales et al., 2008). The former is the process of extracting salient feature vectors from the audio signal, proving a more compact representation of it, while the latter is the process of finding the sequence of words most likely to have generated the salient feature vectors extracted in the previous step.

1.3. The dominance of neural networks

HMM-based ASR systems are still in use today. However, because they make some suboptimal modeling assumptions, they are partially accurate (Tebelskis, 1995). To overcome this limitation, researchers started combining Artificial Neural Networks (ANNs) with HMMs (Trentin & Gori, 2003), as ANNs avoid many of these suboptimal assumptions (Tebelskis, 1995) and can model non-linear functions in a more efficient manner (Nassif et al., 2019).

Nowadays, with the increased availability of computing power and training data, deep neural networks (DNNs) combined with HMMs have become the state of the art in speech recognition (Yu & Deng, 2016). DNNs are ANNs with different levels of non-linear operations (Nassif et al., 2019), allowing to model more complex functions.

Most of the main ASR technologies nowadays use DNN architectures, such as Google, Amazon, Microsoft, or IBM ASR systems.

1.4. The importance of automatic speech recognition

As increasing amounts of spoken content started becoming available, ASR systems gained additional importance, primarily due to the difficulty and inefficiency of accessing and reviewing speech documents (Eskevich & Jones, 2014). That is why voice assistants use ASR to respond to and answer queries (Arya et al., 2020). Castbox (Google Cloud, 2019) is an example of a company that has improved consumer experience by leveraging ASR to allow for more efficient and accurate searches within vast podcast libraries.

However, these systems do not only allow efficient retrieval, review, and reuse of speech documents, but they also allow knowledge extraction from speech (Furui et al., 2004). And this is not only applicable to speech documents but also videos, as we can extract audio from them (Radha, 2016). Because of all these reasons, ASR systems have a wide range of potential applications in marketing and consumer research, which are described below and summarized in Table 1.

ASR systems can be used to automatically transcribe video and audio data related to users and consumers. These transcripts can be coupled with Natural Language Processing (NLP), such as topic modeling or sentiment analysis, or other techniques to extract knowledge from them, e.g., consumer satisfaction, identifying the main topics consumers usually talk about or identifying consumer pain points and areas for improvement.

For instance, in consumer service calls, these transcripts can be used to automate quality monitoring (Zweig et al., 2006), and identify consumer satisfaction (Gavalda & Schlueter, 2010). Intuit (AWS, 2020b), KPMG (Microsoft, 2020b), and HSBC (Google Cloud, 2020) are examples of companies that analyze these ASR transcripts using NLP techniques to extract information about their consumer calls experience, automatically detect areas that require special attention, and highlight opportunities to improve their quality and experience.

In addition, ASR can be used to automatically analyze social media content, such as YouTube or TikTok videos, related to a brand or a product to better understand how consumers perceive and talk about that brand or product (Kaushik et al., 2013; Z. Wang et al., 2022), as well as to monitor brand usage (Damiano & Inguanez, 2019).

Table 1. Summary of main applications of ASR in Marketing

Application	Examples
Efficient retrieval, review, and reuse of speech documents	Castbox (Google Cloud, 2019)
Quality monitoring & satisfaction in consumer service calls	(Gavalda & Schlueter, 2010; Zweig et al., 2006)
Detection of critical areas and improvement opportunities on consumer calls	Intuit (AWS, 2020b), KPMG (Microsoft, 2020b), and HSBC (Google Cloud, 2020)
Brand/Product sentiment and usage on social media	(Damiano & Inguanez, 2019; Kaushik et al., 2013; Z. Wang et al., 2022)

2. Text-to-Speech

In contrast to ASR, TTS is the process of converting text to synthetic speech, intending to produce intelligible and natural-sounding speech (Dutoit, 1997). The main initial application of TTS systems was reading aids for the blind and visually impaired (Lerner, 1982). However, as the technology has progressed, TTS systems have found a wide range of other applications including, but not limited to, speech-enabled devices, virtual assistants, and automated consumer service systems.

2.1. Concatenative synthesis

At first glance, one may intuitively think that TTS could be achieved by storing natural waveforms for each word in a specific language and concatenating them according to the provided text, thus converting them into speech (Klatt, 1987). This basic idea is the foundation for concatenative synthesis, a method that converts text to speech by concatenating prerecorded natural speech samples. The first major study of concatenative speech was published in 1953 by Cyril Harris (Harris, 1953), in which he discussed how segments of speech could be recorded in magnetic tape, which could then be joined together to form words and, consequently speech. Harris's study laid the groundwork for later researchers to build upon, and concatenative synthesis became the dominant way to synthesize speech (Taylor, 2009). Furthermore, early on these systems were recognized for their potential to revolutionize multimedia products such as video games (Lerner, 1982) and movies (Best, 1980), taking the consumer experience to new heights by adding dynamic speech elements.

2.2. From memorizing to learning

Concatenative synthesis is a form of data-driven TTS, in the sense that the synthesis is done by using prerecorded segments of speech. This can be understood as a form of memorization, as these segments are, in fact, directly reused to generate words or phrases (Taylor, 2009). However, memorizing the data reduces the flexibility we can achieve with such a system. For instance, we can't easily convert the original voice into a different one or change the way different words are said (e.g., with different emphasis or intonation).

Another approach would be to use statistical models to learn from data and extract common patterns that makeup speech. This is achieved by systematically analyzing large amounts of labeled speech data to identify the underlying patterns and regularities in the data and, thus, learning the mapping between words and phonemes. This knowledge is then used to generate new, synthetic speech.

Not only does this enable the model to produce the correct speech sounds given a set of written words, but it also allows the model to produce the desired prosody given the context by learning how different features such as intonation, stress, and rhythm are used in a particular sentence to, for example, convey emotionality or ask questions. Prior research has demonstrated that different vocal features greatly influence the listener's perception of the speaker's traits and emotions (Hildebrand et al., 2020; Niculescu et al., 2013), ultimately impacting consumer cognitive responses (Chattopadhyay et al., 2003) and enabling the creation of compelling consumer connections (Sotolongo & Copulsky, 2018).

Therefore, statistical models enable higher flexibility by being able to generate novel utterances, i.e., utterances for which we do not have natural recordings, generate different intonations, or even completely modify the output voice (Taylor, 2009).

Despite all these benefits, generalizing from the limited set of contexts seen in the training data to the almost unlimited number of unseen contexts that will be encountered when synthesizing novel data is not an easy task (S. King, 2011), as this requires a big enough training data set with multiple variations (e.g., in terms of speaker, message, intonation, emotion, context, etc.), which has typically been very difficult and costly to record (Black, 2003).

2.2.1 Hidden Markov Models

As previously seen, HMMs proved to be a successful method of recognizing speech, being extensively used for ASR. A similar approach was developed by Falaschi (Falaschi et al., 1989) to synthesize speech from HMMs, and in just a few years, HMM-based systems were largely used for TTS (Kayte et al., 2015; Tokuda et al., 2000).

In TTS, HMMs are used to generate a sequence of acoustic features, such as the pitch, duration, and spectral parameters (Tóth & Németh, 2010). These parameters together with the phonetic transcription of the text form the output speech signal. Moreover, these parameters can be transformed by using several techniques such as multiple regression, or interpolation to change the voice characteristics, the speaking style, or the intonation of the voice (Zen et al., 2007).

2.2.2 Neural Networks

While HMM-based techniques began to be used for TTS, several researchers also considered using ANN-based techniques for TTS (Karaali et al., 1996; Riedi, 1995; Weijters & Thole, 1993; Xiang & Bi, 1990). However, these early ANN-based techniques did not prove especially successful against HMM-based techniques, which became predominant in TTS (Wu et al., 2016). Recently, following the trend in ASR this has changed, thanks to increased computational power and data availability that have made it possible to train more complex neural networks, i.e., DNNs (Wu et al., 2016). Prior studies support that synthesized speech with DNN-based techniques sounds more natural than HMM-based techniques (e.g., Koriyama & Kobayashi, 2015; Wu et al., 2016; Ze et al., 2013; Zen & Senior, 2014). Furthermore, they provide high adaptability and controllability, as the synthesized voice can be transformed by, for example, using feature space transformations, or augmenting speaker-specific features as input (Wu et al., 2015). Most of the main TTS systems are powered by DNN, such as Google, Amazon, or Microsoft TTS systems.

2.2.3 The importance of Text-to-Speech

The ability to automatically generate speech from text provides significant applications in several fields. For instance, it can be used for disability inclusion, providing a means for blind or low-vision people to receive information presented in text form. It can also be used for language learning, providing a way for language learners to hear how native speakers pronounce words and phrases. Or, as previously said, it provides the means to give a voice to AI-powered voice assistants, enabling them to interact with humans via speech. For instance, Swisscom (Microsoft, 2021) and the BBC (Microsoft, 2020a) have built voice assistants systems to create more engaging consumer experiences. For interested readers, section 3 of the appendix provides a short section on how ASR and TTS systems are leveraged to create conversational interfaces.

Moreover, when paired with ASR techniques, it provides the capability to automate services that require human interaction, such as consumer service or technical support. Several companies such as GE Appliances, Best Western International, and National Australia Bank have implemented TTS technologies to automate several simple tasks in consumer calls, enhancing the overall consumer experience (AWS, 2020a).

This has many consequences for businesses, as it allows them to reduce costs and provide higher interaction with their consumers. For example, such systems can improve consumer service experience (Følstad & Skjuve, 2019; L. Wang et al., 2020), and consumer engagement (Malodia et al., 2021; Moriuchi, 2019).

Furthermore, several media and publishing companies, such as The Washington Post, USA Today, and ProQuest have integrated TTS systems to automatically convert text content into audio (AWS, 2020a). Some media companies, such as CBSi, the South Korean subsidiary of CBS, have even allowed advertisers to create ads using TTS, reducing the costs to create advertisements, and allowing them to easily modify the generated voice (AWS, 2020a).

In addition, the great flexibility provided by TTS systems in changing voice parameters allows businesses to leverage them to, for example, create more persuasive (Zoghaib, 2019), or trustworthy (Elkins & Derrick, 2013) voices.

Table 2. Summary of main applications of TTS in Marketing

Application	**Examples**
AI-powered voice assistants	Swisscom (Microsoft, 2021) and BBC (Microsoft, 2020a)
Automating services that require human interaction	GE Appliances, Best Western International, National Australia Bank (AWS, 2020a)
Creation of advertisements	CBSi (AWS, 2020a)
Automatically converting written media products (news, books, etc.) into audio content	The Washington Post, ProQuest, USA Today (AWS, 2020a)
Create custom voices to change perceived traits, such as persuasiveness or trustworthiness	(Elkins & Derrick, 2013; Zoghaib, 2019)

3 Conclusion

Speech is the primary means of human communication and interaction. Despite this, text was the only means of interacting with computers until recently. With the advancements in ASR and TTS technologies, the ability to recognize and respond to speech has become a reality.

By enabling computers to recognize and respond to speech, these technologies are opening up new avenues for human-computer interaction, such as voice assistants, and offer vast potential for improving consumer experience and engagement.

In this paper, we offer a comprehensive introduction to the capabilities and limitations of ASR and TTS technologies, how they complement each other, and how they can be leveraged to improve customer experience and interaction.

For instance, ASR can be used to efficiently retrieve audio and video content, as well as to analyze what's being said and detect potential focus points. TTS can be used to automatically convert text news and media into audio, making this content suitable for a wider audience at a very low cost. And to create personalized voices that are perceived in a specific way.

ASR and TTS can be combined to, for instance, automate simple tasks in consumer calls, reducing costs and improving customer experience. In addition, with NLP advancements, such as large language models (e.g., GPT), these technologies can be paired to generate conversational interfaces that are more human-like.

All these applications can provide numerous benefits for businesses, such as reducing the need for manual data entry, improving customer service through voice-activated customer service systems, and tracking brand usage.

Overall, ASR and TTS technologies are rapidly changing the landscape of human-computer interaction, and the potential for their use in marketing is enormous. By providing marketers with a comprehensive understanding of these technologies, this paper serves as a valuable resource for marketers looking to leverage these cutting-edge technologies to improve customer experience and engagement. With the continued advancements in these technologies, the future of human-computer interaction is an exciting and promising one.

References

Acero, A. (1992). *Acoustical and environmental robustness in automatic speech recognition* (Vol. 201). Springer Science & Business Media.

Allen, J., Hunnicutt, M. S., Klatt, D. H., Armstrong, R. C., & Pisoni, D. B. (1987). *From text to speech: The MITalk system*. Cambridge University Press.

Arya, S. D., Patel, D., & others. (2020). Implementation of Google Assistant & Amazon Alexa on Raspberry Pi. *ArXiv Preprint ArXiv:2006.08220*.

AWS. (2020a). *Amazon Polly Customers*. AWS.

AWS. (2020b). *Intuit Improves Customer Experience with Amazon Connect, AI-Powered Omnichannel Contact Center*. AWS. https://aws.amazon.com/solutions/case-studies/intuit-contact-center-case-study/

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., & others. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10–11), 763–786.

Best, R. M. (1980). Movies That Talk Back. *IEEE Transactions on Consumer Electronics*, *3*, 670–675.

Black, A. W. (2003). Unit selection and emotional speech. *Interspeech*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2.8457&rep=rep1&type=pdf

Capgemini Digital Transformation Institute. (2018). *Conversational Commerce: Why Consumers Are Embracing Voice Assistants*. Capgemini. https://www.capgemini.com/it-it/resources/conversational-commerce-why-consumers-are-embracing-voice-assistants-in-their-lives/

Chattopadhyay, A., Dahl, D. W., Ritchie, R. J. B., & Shahin, K. N. (2003). Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. *Journal of Consumer Psychology*, *13*(3), 198–204.

Damiano, N., & Inguanez, F. (2019). Brand usage detection via audio streams. *2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin)*, 180–183.

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, *24*(6), 637–642.

de Silva, C. L., & McKnight, C. (1966). *A Case Study of the Feasibility of an Electronic Library Serving a University Research Institute*. Loughborough University.

Denes, P., & Mathews, M. V. (1960). Spoken digit recognition using time-frequency pattern matching. *The Journal of the Acoustical Society of America*, *32*(11), 1450–1455.

Du, J., Wang, Q., Gao, T., Xu, Y., Dai, L.-R., & Lee, C.-H. (2014). Robust speech recognition with speech enhanced deep neural networks. *Fifteenth Annual Conference of the International Speech Communication Association*.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Kluwer Academic Publishers.

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316.

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. *Group Decision and Negotiation*, *22*(5), 897–913.

Eskevich, M., & Jones, G. J. F. (2014). Exploring speech retrieval from meetings using the AMI corpus. *Computer Speech \& Language*, *28*(5), 1021–1044.

Falaschi, A., Giustiniani, M., & Verola, M. (1989). A hidden Markov model approach to speech synthesis. *First European Conference on Speech Communication and Technology*.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399–423.

Flanagan, J. L. (1976). Computers that talk and listen: Man-machine communication by voice. *Proceedings of the IEEE*, *64*(4), 405–415.

Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: user experience and motivation. *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1–9.

Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, *12*(4), 401–408.

Gales, M., Young, S., & others. (2008). The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195–304.

Garg, K., & Jain, G. (2016). A comparative study of noise reduction techniques for automatic speech recognition systems. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2098–2103.

Gavalda, M., & Schlueter, J. (2010). “The Truth is Out There”: Using Advanced Speech Analytics to Learn Why Customers Call Help-line Desks and How Effectively They Are Being Served by the Call Center Agent. In *Advances in Speech Recognition* (pp. 221–243). Springer.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–126.

Google Cloud. (2019). *Castbox: Giving users a voice with Google Cloud*. Google Cloud. https://cloud.google.com/customers/castbox/

Google Cloud. (2020). *HSBC: Using Google Cloud to improve customer service at call centers*. Google Cloud. https://cloud.google.com/customers/hsbc

Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., & Novak, T. P. (2020). Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, *121*, 364–374.

Holmes, J., & Holmes, W. (2002). Speech Synthesis and Recognition. *Speech Synthesis and Recognition*. https://doi.org/10.1201/9781315272702/SPEECH-SYNTHESIS-RECOGNITION-WENDY-HOLMES

Holmgren, J. E. (1983). Human Factors and Behavioral Science: Toward Bell System Applications of Automatic Speech Recognition. *Bell System Technical Journal*, *62*(6), 1865–1880.

Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251–272.

Karaali, O., Corrigan, G., & Gerson, I. (1996). Speech synthesis with neural networks. *World Congress on Neural Networks, San Diego*, 45–50.

Kaushik, L., Sangwan, A., & Hansen, J. H. L. (2013). Sentiment extraction from natural audio streams. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8485–8489.

Kayte, S., Mundada, M., & Gujrathi, J. (2015). Hidden Markov model based speech synthesis: A review. *International Journal of Computer Applications*, *130*(3), 35–39.

Khan, R. A., Scholar, R., Jhunjhunu, J. J. T. U., & Chitode, I. J. S. (2016a). Concatenative speech synthesis: A Review. *Citeseer*, *136*(3), 975–8887. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.742.2786&rep=rep1&type=pdf

Khan, R. A., Scholar, R., Jhunjhunu, J. J. T. U., & Chitode, I. J. S. (2016b). Concatenative speech synthesis: A Review. *Citeseer*, *136*(3), 975–8887. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.742.2786&rep=rep1&type=pdf

King, J. H., & Tunis, C. J. (1966). Some experiments in spoken word recognition. *IBM Journal of Research and Development*, *10*(1), 65–79.

King, S. (2011). An introduction to statistical parametric speech synthesis. *Sadhana 2011 36:5*, *36*(5), 837–852. https://doi.org/10.1007/S12046-011-0048-Y

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, *82*(3), 737–793.

Koriyama, T., & Kobayashi, T. (2015). A comparison of speech synthesis systems based on GPR, HMM, and DNN with a small amount of training data. *Sixteenth Annual Conference of the International Speech Communication Association*.

Lerner, E. J. (1982). Computers: Products that talk: Speech-synthesis devices are being incorporated into dozens of products as difficult technical problems are solved. *IEEE Spectrum*, *19*(7), 32–37.

Malodia, S., Islam, N., Kaur, P., & Dhir, A. (2021). Why do people use Artificial Intelligence (AI)-enabled voice assistants? *IEEE Transactions on Engineering Management*.

Martin, T. B., Nelson, A. L., & Zadell, H. J. (1964). *Speech recognition by feature-abstraction techniques*.

Microsoft. (2020a). *BBC innovates how it delivers trusted news and entertainment with Azure AI*. Microsoft Customer Story. https://customers.microsoft.com/en-us/story/754836-bbc-media-entertainment-azure

Microsoft. (2020b). *KPMG helps financial institutions save millions in compliance costs with Azure Cognitive Services*. Microsoft Customer Stories. https://customers.microsoft.com/en-us/story/754840-kpmg-partner-professional-services-azure

Microsoft. (2021). *How Microsoft brought Swiss voice assistants to life*. Microsoft Customer Stories. https://news.microsoft.com/de-ch/2021/02/05/how-microsoft-brought-swiss-voice-assistants-to-life/

Moriuchi, E. (2019). Okay, Google!: An empirical study on voice assistants on consumer engagement and loyalty. *Psychology & Marketing*, *36*(5), 489–501.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, *7*, 19143–19165.

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *International Journal of Social Robotics*, *5*, 171–191.

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965–2979.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *ArXiv Preprint ArXiv:1904.08779*.

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *Ieee Assp Magazine*, *3*(1), 4–16.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Radha, N. (2016). Video retrieval using speech and text in video. *2016 International Conference on Inventive Computation Technologies (ICICT)*, *2*, 1–6.

Riedi, M. (1995). A neural-network-based model of segmental duration for speech synthesis. *Fourth European Conference on Speech Communication and Technology*.

Sotolongo, N., & Copulsky, J. (2018). Conversational marketing: Creating compelling customer connections. *Applied Marketing Analytics*, *4*(1), 6–21.

Tabet, Y., & Boughazi, M. (2011). Speech synthesis techniques. A survey. *7th International Workshop on Systems, Signal Processing and Their Applications, WoSSPA 2011*, 67–70. https://doi.org/10.1109/WOSSPA.2011.5931414

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press. https://books.google.com/books?hl=en&lr=&id=BFnkm-FpBAUC&oi=fnd&pg=PR9&dq=text-to-speech+synthesis+taylor&ots=udr6kRQ1aT&sig=I7LKLumrdQhXos2mmVzr6BpwtnY

Tebelskis, J. M. (1995). *Speech recognition using neural networks*. Carnegie Mellon University.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, *3*, 1315–1318.

Tóth, B., & Németh, G. (2010). Improvements of Hungarian hidden Markov model-based text-to-speech synthesis. *Acta Cybernetica*, *19*(4), 715–731.

Trentin, E., & Gori, M. (2003). Robust combination of neural networks and hidden Markov models for speech recognition. *IEEE Transactions on Neural Networks*, *14*(6), 1519–1531. https://doi.org/10.1109/TNN.2003.820838

Vonusa, R. S., Jacobs, L., & Beek, B. (1970). *An Automatic Isolated Word Recognition System Study and Evaluation.*

Waibel, A., & Lee, K.-F. (1990). *Readings in speech recognition*. Morgan Kaufmann.

Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2020). Voice-Based AI in Call Center Customer Service: Evidence from a Field Experiment. *Call Center Customer Service: Evidence from a Field Experiment (September 1, 2020)*.

Wang, Z., Gao, P., & Chu, X. (2022). Sentiment analysis from Customer-generated online videos on product review using topic modeling and Multi-attention BLSTM. *Advanced Engineering Informatics*, *52*, 101588.

Washani, N., & Sharma, S. (2015). Speech recognition system: A review. *International Journal of Computer Applications*, *115*(18), 7–10.

Wattenbarger, B. L., Garberg, R. B., Halpern, E. S., & Lively, B. L. (1993). Serving Customers With Automatic Speech Recognition—Human-Factors Issues. *AT&T Technical Journal*, *72*(3), 28–41.

Weijters, T., & Thole, J. (1993). Speech synthesis with artificial neural networks. *IEEE International Conference on Neural Networks*, 1764–1769.

Wu, Z., Swietojanski, P., Veaux, C., Renals, S., & King, S. (2015). A study of speaker adaptation for DNN-based speech synthesis. *Sixteenth Annual Conference of the International Speech Communication Association*.

Wu, Z., Watts, O., & King, S. (2016). Merlin: An Open Source Neural Network Speech Synthesis System. *9th ISCA Speech Synthesis Workshop*, 202–207. https://doi.org/10.21437/SSW.2016-33

Xiang, Z., & Bi, G. (1990). A neural network model for Chinese speech synthesis. *1990 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1859–1862.

Yu, D., & Deng, L. (2016). *Automatic speech recognition* (Vol. 1). Springer.

Ze, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. *2013 Ieee International Conference on Acoustics, Speech and Signal Processing*, 7962–7966.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In P. Wagner, J. Abresch, S. Breuer, & W. Hess (Eds.), *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)* (pp. 294–299). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.69.5302&rep=rep1&type=pdf

Zen, H., & Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3844–3848.

Zoghaib, A. (2019). Persuasion of voices: The effects of a speaker's voice characteristics and gender on consumers' responses. *Recherche et Applications En Marketing (English Edition)*, *34*(3), 83–110.

Zweig, G., Siohan, O., Saon, G., Ramabhadran, B., Povey, D., Mangu, L., & Kingsbury, B. (2006). Automated quality monitoring for call centers using speech and NLP technologies. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, 292–295.

Online Supplement

This online supplement lists a series of resources that may be useful for researchers interested in learning more about Automatic Speech Recognition (ASR) and Text-to-Speech (TTS). Lastly, it briefly explains how ASR and TTS can be combined to create voice bots.

1. Text-to-Speech

1.1 Text-to-Speech Theory

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis* (Vol. 3). Springer Science & Business Media.

1.2 Main Text-to-Speech APIs

- Google: Text-to-Speech: Lifelike Speech Synthesis | Google Cloud
- IBM: IBM Watson Text to Speech | IBM
- Amazon: Text to Speech Software – Amazon Polly – Amazon Web Services
- Microsoft: Text to Speech – Realistic AI Voice Generator | Microsoft Azure

Each of these APIs provides comprehensive information and tutorials on how to use them.

1.3 Text-to-Speech Courses

- It Speaks! Create Synthetic Speech Using Text-to-Speech

  Course offered by Google, which showcases how to use their TTS API

  https://www.coursera.org/projects/googlecloud-it-speaks-create-synthetic-speech-using-text-to-speech-7wx56
- Recognize and synthesize speech

  Course offered by Microsoft, which showcases how to use speech synthesis and speech recognition using the Microsoft APIs:

  https://learn.microsoft.com/en-us/training/modules/recognize-synthesize-speech/

- Watson Speech to Text Service V2

  Course offered by IBM, which showcases how to use their TTS API:

  https://www.ibm.com/training/badge/watson-speech-to-text

- Build a Serverless Text-to-Speech Application with Amazon Polly

  Course offered by Amazon, which showcases how to use their TTS API:

  https://explore.skillbuilder.aws/learn/course/external/view/elearning/1094/build-a-serverless-text-to-speech-application-with-amazon-polly

2. Automatic Speech Recognition Theory

2.1 Automatic Speech Recognition Theory

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep learning for NLP and speech recognition* (Vol. 84). Cham, Switzerland: Springer.

Yu, D., & Deng, L. (2016). Automatic speech recognition (Vol. 1). Berlin: Springer.

Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, 67.

2.2 Main Automatic Speech Recognition APIs

- Google: Speech-to-Text: Automatic Speech Recognition | Google Cloud
- IBM: IBM Watson Speech to Text | IBM
- Amazon: Amazon Transcribe – Speech to Text - AWS
- Microsoft: Speech to Text – Audio to Text Translation | Microsoft Azure

Each of these APIs provides comprehensive information and tutorials on how to use them.

2.3 Automatic Speech Recognition Courses

- Build automated speech systems with Azure Cognitive Services

  Course offered by Microsoft, which showcases how to use Microsoft Automatic Speech Recognition API.

https://www.coursera.org/projects/build-automated-speech-systems-with-azure-cognitive-services

- Introduction to Amazon Transcribe

  Course offered by Amazon, which showcases how to use Amazon Automatic Speech Recognition API.

  https://explore.skillbuilder.aws/learn/course/external/view/elearning/224/introduction-to-amazon-transcribe

- Speech to Text Transcription with the Cloud Speech API

  Course offered by Amazon, which showcases how to use Google Automatic Speech Recognition API.

  https://www.coursera.org/projects/googlecloud-speech-to-text-transcription-with-the-cloud-speech-api-vwhbe

- Getting started with Speech to Text

  Short documentation explaining how to use IBM Automatic Speech Recognition API.

  https://cloud.ibm.com/docs/speech-to-text?topic=speech-to-text-gettingStarted

3. Integrating ASR and TTS to generate voice bots

Most voice bots typically consist of five key components: a recording device, a speaker, an ASR system, a TTS system, and a dialogue system. The recording device and speaker function as the input and output devices, respectively. The ASR system translates the user's speech into text, while the TTS system converts the bot's response into spoken language. The dialogue system acts as the bot's "brain", processing and interpreting the user's input and generating an appropriate response.

When a user asks a question to a voice bot, the recording device captures the user's speech and sends it to the ASR system, which transcribes it into text. The text is then sent to the dialogue system, which processes and interprets it to determine the appropriate response. Once

the response is generated, it is sent to the TTS system, which converts it into spoken language that is played through the speaker to the user. This entire process happens in a matter of seconds, allowing for a seamless conversation between the user and the voice bot. Figure 1 visually depicts this process.

For readers interested in further details about this process or on what types of dialogue systems exist and how they work, we recommend *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots* by Michael McTear (Ulster University), doi:10.2200/S01060ED1V01Y202010HLT048.

Fig. 1 Schematic and simplified representation of how a voice bot operate