Utterance Selection for Optimizing Intelligibility of TTS Voices Trained on ASR Data

Erica Cooper[1]*, Xinyue Wang*[1]*, Alison Chang*[2]*, Yocheved Levitan*[1]*, Julia Hirschberg*[1]

[1]Columbia University, USA
[2]Google, USA

ecooper@cs.columbia.edu, xw2368@columbia.edu, alisonchang@google.com,
yocheved.levitan@gmail.com, julia@cs.columbia.edu

Abstract

This paper describes experiments in training HMM-based text-to-speech (TTS) voices on data collected for Automatic Speech Recognition (ASR) training. We compare a number of filtering techniques designed to identify the best utterances from a noisy, multi-speaker corpus for training voices, to exclude speech containing noise and to include speech close in nature to more traditionally-collected TTS corpora. We also evaluate the use of automatic speech recognizers for intelligibility assessment in comparison with crowdsourcing methods. While the goal of this work is to develop natural-sounding and intelligible TTS voices in Low Resource Languages (LRLs) rapidly and easily, without the expense of recording data specifically for this purpose, we focus on English initially to identify the best filtering techniques and evaluation methods. We find that, when a large amount of data is available, selecting from the corpus based on criteria such as standard deviation of f0, fast speaking rate, and hypo-articulation produces the most intelligible voices.

Index Terms: speech synthesis, parametric synthesis, data selection, found data, crowdsourcing.

1. Introduction

Speech technology has progressed enormously during the past decade, with widespread adoption of spoken dialogue systems like Siri, Cortana, and the Echo. There now exists high quality text-to-speech (TTS) synthesis for High Resource Languages (HRLs) such as English, German, Mandarin, Japanese, and Spanish – these languages are fortunate to have expert-built pronunciation rules and dictionaries, part-of-speech taggers, and language models, as well as large amounts of data from professional voice talents for developing state-of-the-art TTS systems. However, this is not the case for every language: There are approximately 6500 languages in the world, many spoken by millions of people, which have no such resources. LRLs such as Telugu, Mongolian, Vietnamese, Cebuano, and Amharic, for example, have few natural language processing resources available and have been little studied for TTS. Thus, speakers of these languages are deprived of speech-related technologies that allow communication with devices for those without reading skills or across language barriers in applications such as spoken dialogue systems or speech-to-speech translation systems. Our goal is to develop techniques to produce TTS systems for such languages easily by maximizing the utility of existing sources of data which have been created for other purposes, choosing subsets and filtering appropriately to produce material suitable for building intelligible and natural TTS systems.

In the LRL setting, we do not have access to a large corpus of high-quality, single-speaker data, as this is very expensive and time-consuming to collect and thus typically requires a major economic motivation. However, for many LRLs, there are large amounts of "found" data that can be acquired cheaply and easily (as on mobile phones or by web scraping), or that have already been collected for other purposes, such as ASR. While recording conditions, and thus data quality, do not approach the standards usually required to build TTS systems, the development of Hidden Markov Model (HMM) and other forms of parametric speech synthesis [1] have made it possible to train TTS systems on heterogeneous data. Although there has been some prior work on training parametric synthesizers with found data, there has not been a systematic evaluation of the different types of found data that can be used to produce voices, or of methods for producing the most natural and intelligible voices from these sources.

This paper describes research on filtering techniques needed to produce intelligible TTS voices from noisy, multi-speaker ASR data. We explore methods of data selection on the ASR data which can be used to identify the best utterances for voice training in a corpus, while excluding utterances that introduce excessive noise or artifacts. While our ultimate goal is to facilitate the rapid development of natural-sounding and intelligible TTS voices in LRL corpora, we first evaluate our methods on American English telephone speech, to facilitate quicker experimentation and evaluation. We use crowdsourcing to evaluate intelligibility and compare human transcription-based intelligibility evaluation to performance of several ASR APIs to speed the evaluation process.

2. Related Work

Other work on the use of "found" data for building TTS voices has often involved adaptation of voices trained on clean data with noisier recordings. In [2], noisy recordings of political speeches were used to adapt an average HMM voice trained on clean data from many speakers. The authors obtained a robust, natural-sounding voice with performance minimally degraded by the inclusion of noisy data. [3] trained an average voice on data collected in an office environment and adapted to cleanly-recorded speech. They found that using both noisy and clean data together produced a voice with a slightly (but not statistically-significantly) higher mean opinion score (MOS) than a voice trained on clean data alone, and concluded that more data, even of a lesser quality, can be beneficial. [4] also trained average voices on clean data from a TTS corpus and adapted it using data with added noise. They found that listeners could distinguish between voices adapted with clean and noisy data, but that naturalness and speaker similarity were not affected. [5] used radio broadcast news recordings to train voices, investigating different speaker diarization and background mu-

sic and noise detection techniques to remove noisy utterances automatically. Finally, audiobooks have been a popular source of "found" data for building TTS voices due to their clean recording conditions and the fact that they typically contain large amounts of speech from a single speaker [6] [7] [8].

While these researchers have begun to investigate the use of "found" data for TTS synthesis, many questions remain, including the best types and combinations of data to use and the best ways to filter data from sources such as ASR corpora. In this paper we present evaluations of different filtering techniques at the utterance level and their effect on TTS intelligibilty.

3. Characteristics of a "Good" TTS Voice

In previous work [9] [10], we trained HMM voices on radio broadcast news speech from the Boston University Radio News Corpus (BURNC) [11]. We selected subsets of utterances based on a number of different factors we hypothesized might be useful in optimizing for naturalness, such as speaking rate, f0 and energy mean and standard deviation, and level of articulation. We were guided in choosing our features by the instructions that are typically given to voice talents when recording audio data for a unit selection voice. TTS speakers are usually professional voice talents who are instructed to speak as clearly and consistently as possible, without varying their voice quality, speaking style, pitch, volume, or tempo significantly [12]. While these requirements are critical for creating a concatenative or unit-selection voice, in which recorded audio is segmented and then joined with other pieces of the recording, we propose that these constraints will also lead to more consistent models for a parametric voice built from found data. We therefore look at prosodic characteristics such as mean and standard deviation of fundamental frequency (f0), energy, and speaking rate as well as presence of disfluencies and transcribed noise in selecting training material. We have also found in our previous work using political speeches, ads, and interviews that certain ranges of some of these types of features correspond to speech rated as charismatic across multiple cultures [13] [14] [15] [16], leading us to hypothesize that selecting training utterances based on these features may also produce voices that are preferred by listeners. When applying this knowledge to utterance selection for training voices on broadcast news data, we found that hyper-articulated and slow speaking rate utterances produced the least natural-sounding voices. We also discovered that removing utterances that are outliers with respect to hyper-articulation, as well as combining the selection of hypo-articulated utterances and low mean f0 utterances, produced the most natural voices.

4. Corpora and Tools

In our current work, we train voices on data selected from the MACROPHONE [17] corpus, which was designed for the development of telephone-based dialogue systems such as travel booking and other database-related tasks. The utterances were read by 5,000 speakers over the phone. The data contains 83 hours and 31 minutes of data from adult female speakers, 63 hours and 10 minutes from adult male speakers, 1 hour and 33 minutes from adult speakers of unknown gender, 5 hours and 52 minutes from female children, 6 hours and 52 minutes from male children, and 1 hour and 16 minutes from children of unknown gender. This is a representative ASR training corpus of the sort that has begun to become available for LRLs.

We trained our TTS voices using the Hidden Markov Model Based Speech Synthesis System (HTS) [18]. We used the speaker-independent (SI) training recipe for HTS version 2.3. In our prior work [9], we found that using speaker-adaptive training (SAT) did not give a significant improvement in naturalness for voices trained on BURNC. While there is the possibility that using SAT may improve intelligibility, the time and computational resources required for SAT on the large number of speakers represented in MACROPHONE would be prohibitive to our goal of rapid experimental iteration, so in this work all of our voices are trained speaker-independently. We will examine the benefits of SAT on our most highly rated voices in future work. We obtain the standard set of full-context phonetic labels using the Festival Speech Synthesis System front-end [19]. Synthesis and vocoding were done using hts-engine.

5. Crowdsourced Evaluation of Intelligibility

To evaluate the intelligibility of our voices, we published crowdsourced listening tests online using Amazon Mechanical Turk (MTurk), a popular crowdsourcing platform. To restrict our listeners to native speakers of English, we required all workers to complete a qualification test in which workers choose the languages they have spoken since birth from a list of options before attempting any of our tasks. We only allowed workers who selected English as one of these languages to participate in our evaluation and also excluded workers who chose more than three languages in total in order to safeguard against those who might select many languages in an attempt simply to pass the test. We also restricted our tasks' visibility to workers within the United States.

We produced 10 syntactically-sound but semantically unpredictable sentences (SUS) of the standard form `det adj noun verb det adj noun` as used in the Blizzard challenge [20], and synthesized them with each of the voices described below. We also included one semantically-predictable sentence spoken clearly as an attention check question. This is a standard intelligibility test for TTS which has been shown to be viable for crowdsourcing [21]. Workers were asked to transcribe each of the eleven sentences, presented in random order. Since the sentences were the same for each voice, to enable a sensible comparison across voices, workers were only allowed to transcribe sentences for one voice, to remove any bias arising from the workers remembering the sentences. Five workers transcribed all sentences for each voice.

After sentences were transcribed, we computed word error rate (WER) for each voice (averaging over each of the five workers) to measure intelligibility, comparing results to the text that was actually synthesized. Since the transcriptions were typed by humans, they were prone to typographical errors and misspellings, which we hand-corrected. We also allowed singular/plural confusions, such as "musical" / "musicals," but we did not allow confusions between words with the same stem, such as "fragrant" / "fragrance." We also allowed compound word variants, such as "blackbird" / "black bird."

We also explored the use of automatic speech recognition (ASR) as an alternative method of evaluating for intelligibility. This is further described in Section 7.

6. Filtering Techniques

We began with the first 10 hours of utterances labeled as being spoken by adult female speakers in the MACROPHONE corpus, and then extended it to the entire 83 hours of female speech, to compare the effects of having a smaller or larger pool of data from which to select. We selected 2-hour subsets from just the

first 10 hours of data, and then 2- and 4-hour subsets from the full 83 hours. We selected our training subsets based on criteria such as mean and standard deviation of f0 and energy, as well as speaking rate (computed as syllables per second), level of articulation (computed as mean energy divided by speaking rate), and utterance length. For each feature, we computed its value for each utterance, and then sorted the list of utterances from low to high. Then, we obtained subsets by selecting e.g. the first two hours' worth of utterances from that list. We also experimented with removing different types of utterances that we hypothesized might hurt the quality of the voice, such as very short utterances of only one or two words; utterances containing clipped audio; utterances containing transcribed noise (such as "[unintelligible]," "[bg_speech]," and "[line_noise]"); and utterances consisting of a word spelled out letter-by-letter, which are indicated in the corpus by "spword" in the file name.

We trained our baseline voice on all of the first 10 hours of female utterances since training on the full 83 hours would be prohibitively computationally expensive. This produced a voice with a word error rate of **67.7%** when transcribed by MTurk workers. We compare all of our MACROPHONE subset voices to this baseline. Results are shown in Tables 1 and 2. Voices that did better than the 10-hour baseline appear in bold.

Table 1: *Word error rates for voices trained on 2-hour subsets of the first 10 hours of female MACROPHONE data, selected based on prosodic features.*

Feature	Low	Med	High
Mean f0	98.6	85.7	100.3
Stdv f0	83.1	80.0	87.1
Mean energy	98.6	95.7	70.6
Stdv energy	100.9	85.4	79.7
Speaking rate	-	99.1	**54.3**
Articulation	76.0	87.7	-
Utterance length	96.6	85.4	96.9

Training on the two hours of slowest speaking rate utterances and most hyper-articulated utterances both failed due to lack of phonetic coverage.

Table 2: *Word error rates for voices trained on subsets of the 10 hours of data minus utterances removed based on different noise criteria.*

Subset	Hours	WER
Baseline	*10:00*	*67.7*
3 or more words	7:34	79.7
No clipping	9:57	77.7
No transcribed noise	5:53	**58.9**
No spelled words	9:24	94.3

While the baseline MACROPHONE voice is not rated as highly intelligible, the amount of room for improvement allows us to see approaches that do well — namely, the two-hour subset of utterances with the fastest speaking rate and the subset that excludes utterances with transcribed noise that do beat the MACROPHONE baseline. Removing utterances containing transcribed noise did improve intelligibility, but surprisingly, removing the shortest utterances, utterances containing clipping, and utterances containing spelled-out words did not.

Next, we extended some experiments for our five best selection approaches so far – high mean energy, high standard deviation of energy, fast speaking rate, low articulation level, and middle standard deviation of f0. Rather than limiting ourselves to selecting from just the first 10 hours of data, we selected 2- and 4-hour training subsets from the entire 83 hours of data, minus any utterances containing transcribed noise (since we have shown already that this removal improves synthesis output), to observe whether there is an improvement from having a larger pool of data to select from. Results are shown in Table 3.

Table 3: *Word error rates for voices trained on 2- and 4-hour subsets selected from the full 83 hours of data.*

Feature	2hrs	4hrs
High mean energy	**60.0**	**48.3**
High stdv energy	83.1	**64.6**
Fast speaking rate	**66.6**	**48.3**
Hypo-articulation	**64.6**	**49.1**
Middle stdv f0	**48.0**	**45.1**

When looking at results for selecting 2-hour subsets from just the first 10 hours of data (Table 1), versus selecting 2-hour subsets from the full data set (Table 3, first column), we noticed that selecting subsets from the full dataset does usually produce more intelligible voices than selecting just from the first 10 hours, with most of these voices being rated as more intelligible than the 10-hour baseline, despite being trained on only 1/5 of the amount of data. Extending to 4-hour subsets consistently produces better voices than the baseline, as well. This indicates that more data is only better if it is chosen in a principled way, and validates our hypothesis that better voices can be trained by identifying the best training utterances in a noisy corpus, even if this results in less training data.

Three of our best voices were created from fast speaking rate utterances – both the 4-hour and 2-hour sets selected from the full data set, and the 2-hour set selected from just the first 10 hours. A low level of articulation, which encodes fast speaking rate, also proved to be a preferable feature – voices trained on the 4 hours and 2 hours of most hypo-articulated utterances selected from the full data set scored better than the baseline. Our top two voices were based on selecting utterances with middle values for standard deviation of f0, with the 2-hour subset of the full data at 48% and the 4-hour subset at 45.1%. This was surprising as we would expect *low* values of f0 standard deviation to be more consistent with the speaking style in a standard text-to-speech corpus; however, these corpora are optimizing for naturalness, generally in a unit selection setting, whereas we are optimizing for intelligibility.

7. Automatic Intelligibility Evaluation

A limitation of our approach is the long turnaround time for crowdsourcing voice transcriptions. Since each worker is allowed to transcribe only a single voice, evaluation proceeds slowly regardless of individual workers' interest in the task. This led us to investigate the possibility of using automatic speech recognition (ASR) to evaluate intelligibility. Although an ASR system will not interpret a voice exactly as a human would, depending heavily upon the type of data on which it was trained, it would nevertheless return results very quickly and not have the limitation of remembering and being influenced by repeat sentences. We therefore thought it worthwhile to see how this type of evaluation compares to that done by humans and whether in fact there are some reliable correlations.

We tested three different general-purpose, state-of-the-art, industry-level ASR APIs (Application Programming Interfaces) to determine the viability of their use for evaluating voices:

wit.ai [22], a natural language API toolkit owned by Facebook; Watson [23], IBM's API for cognitive applications; and the Google Cloud Speech API [24]. We decided to try APIs rather than building our own ASR because using state-of-the-art recognizers should presumably provide the best possible proxy for a human listener. Our hypothesis was that some of these recognizers might correlate well with human transcription performance, so we can use these as a first step to choosing our best candidate voices to send to MTurk.

For each voice, we ran the same set of synthesized SUS that we gave to MTurk workers through each ASR API. We then computed WER from the returned transcripts. We allowed the same singular/plural and compound word confusions that we allowed in the transcriptions from Mechanical Turk, but we did not need to correct for spelling or typographical errors.

7.1. Results

We found strong correlations between our three different ASR APIs' WERs and those from MTurk. We report correlations across our 34 different voices in Table 4. Furthermore, for all voices that humans rated as better than baseline, all three ASRs agreed that these voices were better than the baseline. This indicates that using ASR APIs is a promising pre-selection approach to decide which voices should get evaluated on MTurk.

While using ASR is much faster than crowdsourcing, it comes with its own challenges. For example, we noticed that sending the same audio clip multiple times to the same ASR did not necessarily always return the same transcription. A major downside of using an ASR API is that we have no information on the internal system — not just the type of models they are using, but how often they are updated, or whether some machines in their cloud are running different versions of the recognizer. So we can only speculate as to why repeated recognition of the same audio file would result in multiple different 1-best transcripts each time. We originally thought that using an ASR API would serve as a very consistent way of evaluating the voices, but this may not be the case.

Nevertheless, the human evaluations are also somewhat inconsistent and, in fact, they tend to be more inconsistent than the ASR evaluations. We have measured standard deviations in word error rate for the baseline MACROPHONE voice across the 5 workers who transcribed it, as well as standard deviations for those same utterances sent 5 times each to our three ASR APIs, as a way to measure variability of the different systems; these are also reported in Table 4.

Table 4: *Correlation of ASR APIs with MTurk on 34 voices, and standard deviation in WER when evaluating the baseline MACROPHONE voice 5 times.*

Evaluation	Correlation (r)	Std.Dev (%)
MTurk	—	4.52
wit.ai	0.728	1.20
Watson	0.797	0.00
Google	0.876	0.00

Both Watson and Google returned the same transcripts all five times, indicating that they have the least variability.

We have also found challenges related to task specification. While we were able to tell MTurk workers in the instructions that the sentences they were transcribing would not necessarily make sense, we could give no such instructions to an ASR. We noticed that, for example, wit.ai appeared to be attempting to recognize sentences or parts of sentences that "made sense" – most likely because its language model was trained on semantically-predictable data. While wit.ai had the most obvious language model effects, this also applies to *any* ASR API over which we do not have control. Ideally we would be able to "tell" an ASR that the sentences will not necessarily make sense by specifying a very simple language model such as a unigram or bigram, but unfortunately we have no such control over a cloud-based ASR API. Thus, for future work, we plan to see whether using ASR systems trained on the same data as the TTS voices correlates with human judgments, since in the case of actual LRLs, this may be all we have available. This will allow us complete control over the language model and over the system in general so that we can ensure consistent evaluation.

For future work, we also plan to compare more traditional objective measures. Objective measures for intelligibility are typically used for measuring signal loss of natural speech in noise environments or over a noisy transmission line and are often only applicable in limited circumstances [25]; it is rare for these measures to be used for evaluating synthetic speech. However, [26] explored the use of a variety of such metrics to evaluate speech from a state-of-the-art HMM synthesizer under a number of additive noise conditions, finding that some measures correlated well with human intelligibility ratings. While a voice trained on high-quality, single-speaker, TTS-specific data and played in noisy environments is likely to have different intelligibility issues than a voice originally trained on noisy data, it may nevertheless be worthwhile to see whether these measures correlate with human judgments for our voices as well.

8. Conclusions and Future Work

We have found that selecting for utterances with fast speaking rate and removing utterances with transcribed noise can produce an improved TTS voice when only noisy "found" data from multiple speakers is available. We have also found that level of articulation, mean and standard deviation of energy, and standard deviation of f0 are useful selection features. We are currently exploring additional filters and combinations of filters. We also plan to explore different modeling approaches such as adaptation and neural network based synthesis. Using ASR for intelligibility evaluation has shown promise for reducing our experimental iteration time by identifying voices that are the best candidates for human evaluation. Finally, we have begun using these approaches on LRLs (currently, Amharic, Telugu, and Turkish) in addition to English corpora to see how well filters that suit one language generalize to others.

9. Acknowledgements

This work was supported by NSF 1539087 "EAGER: Creating Speech Synthesizers for Low Resource Languages" and by Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. We also thank Meredith Brown for her helpful advice regarding setting up experiments and interpreting results on Mechanical Turk.

10. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] J. Yamagishi, Z. Lin, and S. King, "Robustness of HMM-based speech synthesis," *INTERSPEECH*, 2008.

[3] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis analysis and application of TTS systems built on various ASR corpora," *IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 5*, 2010.

[4] R. Karhila, U. Remes, and M. Kurimo, "HMM-based speech synthesis adaptation using noisy data: Analysis and evaluation methods," *Acoustics, Speech, and Signal Processing*, pp. 6930–6934, 2013.

[5] A. Gallardo-Antolín, J. Montero, and S. King, "A comparison of open-source segmentation architectures for dealing with imperfect data from the media in speech synthesis," *INTERSPEECH*, 2004.

[6] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis, "Using audio books for training a text-to-speech system," *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014.

[7] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "TUNDRA: A multilingual corpus of found data for TTS research created with light supervision," *INTERSPEECH*, 2013.

[8] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," *INTERSPEECH*, 2011.

[9] E. Cooper, Y. Levitan, and J. Hirschberg, "Data selection for naturalness in HMM-based speech synthesis," *Speech Prosody*, 2016.

[10] E. Cooper, A. Chang, Y. Levitan, and J. Hirschberg, "Data selection and adaptation for naturalness in HMM-based speech synthesis," *Interspeech*, 2016.

[11] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Tech. Rep.*, 1995.

[12] J. Matoušek, D. Tihelka, and J. Romportl, "Building of a speech corpus optimised for unit selection tts synthesis," *LREC*, 2008.

[13] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," *Eurospeech*, 2005.

[14] F. Biadsy, J. Hirschberg, A. Rosenberg, and W. Dakka, "Comparing American and Palestinian perceptions of charisma using acoustic-prosodic and lexical analysis," *INTERSPEECH*, 2007.

[15] F. Biadsy, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert, "A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech," *Speech Prosody*, 2008.

[16] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, 2008.

[17] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: An American English telephone speech corpus for the POLYPHONE project," *ICASSP*, 1994.

[18] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," *6th ISCA Workshop on Speech Synthesis*, 2007.

[19] A. W. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system." [Online]. Available: http://www.festvox.org/festival/

[20] A. Black and K. Tokuda, "The blizzard challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," *Proc. Interspeech*, pp. 77–80, 2005.

[21] S. Buchholz, J. Latorre, and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," in *Crowdsourcing for Speech Processing: Applications to Data, Collection, Transcription and Assessment*, M. Eskénazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, Eds. Chichester: John Wiley & Sons, Ltd, 2013, ch. 7, pp. 173–214.

[22] wit.ai: Natural language for developers. [Online]. Available: https://wit.ai/. Accessed March 1, 2017.

[23] Watson developer cloud speech to text. [Online]. Available: https://www.ibm.com/watson/developercloud/speech-to-text.html. Accessed March 1, 2017.

[24] Google cloud speech api (beta). [Online]. Available: https://cloud.google.com/speech/. Accessed March 1, 2017.

[25] A. Schmidt-Nielsen, "Intelligibility and acceptability testing for speech technology," *No. NRL/FR/5530-92-9379. Naval Research Lab*, 1992.

[26] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Evaluation of objective measures for intelligibility prediction of hmm-based synthetic speech in noise," *ICASSP*, 2011.