# Hypo and Hyperarticulated Speech Data Augmentation for Spontaneous Speech Recognition

Sung Joo Lee, Byung-Ok Kang, Hoon Chung, Jeon Gue Park, Yun Keun Lee
Electronics and Telecommunications Research Institute, Daejeon, The Republic of Korea. Email: lee1862@etri.re.kr

***Abstract*—Among many challenges in spontaneous speech recognition, we focus on the variability of speech depending on the degree of articulation such as hypo and hyperarticulation. In this paper, we investigate the feasibility of the past acoustic-phonetic studies on the variability of speech in terms of the data augmentation of a spontaneous speech recognition system. To do so, we develop data augmentation approaches to reflect the acoustic-phonetic characteristics of hypo and hyper-articulated speech. Since our approaches are based on signal processing methods they do not require a model learned from supervised or unsupervised data. A series of speech recognition tests are conducted across various speech styles. The results show that we are able to achieve meaningful performance gain by using our approaches. It also indicates that the past acoustic-phonetic knowledge of the variability of speech is useful for improving the recognition performance of spontaneous speech including hypo and hyper-articulated speech.**

***Keywords—Speech recognition, data augmentation, hypo and hyperarticulation, speech sythesis.***

## I. INTRODUCTION

The state of the art speech recognition technology makes it possible to convert speech into text with high accuracy. This listening competence is essential for today's voice-enabled applications such as Amazon Echo, Google assistant, Microsoft Cortana and so on. However, the recognition performance in spontaneous speech seems still insufficient for user requirement [1], [2]. This is due to the fact that a speech recognition system is mainly built using read speech while the properties of read speech are different from spontaneous speech in the aspect of acoustics and linguistics. Furthermore, it is an expensive and time-consuming work to collect natural speech in the real world. Therefore, it is difficult to create an acoustic model fully representing the distributions of various speech sounds.

In particular, there are typical hurdles in recognizing spontaneous speech. One of them is the so-called disfluency (e.g. filled pauses, hesitations, repetitions, false starts, and word fragments). The formal word sequence of a utterance is disrupted where speech disfluency takes place. This kind of irregularity among the ordinary speech flow makes it difficult to automatically recognize speech. In addition, filled pause and word fragment make the matter worse by producing word-like elements, which are not listed in the lexicon of a speech recognition system. According to the literature in [3], the average value of the ratio of filled pauses and word fragments is over 10% in a spontaneous speech style such as free dialogue. Therefore, speech disfluency has the potential to significantly deteriorate the performance of a speech recognition system. And this disfluency issue has attracted much research interest in the speech recognition research area and many academic activities have been dedicated to the problem [4], [5]. However, this issue still remains in challenging area. Another obstacle in spontaneous speech recognition is that people pronounce diversely across different speech styles. And these acoustic-phonetic variations caused by a human are known as highly complicated phenomena. In the current automatic speech recognition (ASR) research, the most technologies rely on the typical pattern matching paradigm from short-term acoustic observations to most likely phonemic or phoneme-like categories. This mapping task becomes difficult as the acoustic-phonetic variability of speech is increased. Therefore, speech variability is another reason why ASR performance is corrupted. Despite the past research efforts, we don't have much understanding of the structural attributes of speech variability, especially in spontaneous speaking style.

Anyway, it is obvious that different speech styles lead to the increment of speech variability. For example, the relaxed articulatory movement often results in a reduction or skip of the pronunciation of certain phonemes, or syllables. On the other hand, people often enunciate longer and louder on the phrase which is important in their conversational situation. The former speaking style is the so-called hypoarticulation and the latter is categorized into hyperarticulation. Many analytical and technical studies have been conducted on these articulation styles in the past decades. Among them, the acoustic-phonetic features of hypo and hyperarticulated speech in [6] are well explained in the aspect of a speech synthesizer. In this paper, we explore if these characteristics are also effective in natural speech recognition. To do so, we develop data augmentation approaches mimicking the acoustic-phonetic attributes of hypo and hyperarticulated speech. A series of speech recognition tests across several speaking styles (interview, debate, academic presentation and Lombard speech) are conducted to assess the performance of the proposed approaches. The result shows that our approaches are useful for recognizing spontaneous speech as well as distant Lombard speech.

The remainder of this paper is organized as follows. After the introduction in Section I, the proposed speech synthesis approaches are explained in Section II. In Section III, the important points in the implementation of our approaches are described. And then, the procedure and

ISBN 978-90-827970-1-5 

result of evaluating the proposed approaches are provided in Section IV. Finally, some concluding remarks are given in Section V.

## II. Speech Data Augmentation

Although spontaneous speech is common in human communication and may make people feel more comfortable in man-machine interface, our knowledge of spontaneous speech is insufficient to achieve the necessary breakthrough in ASR. Therefore, spontaneous speech recognition has received much research attention in the last decades. Among the difficulties in identifying spontaneous speech into text by computers, we are particularly interested in the acoustic-phonetic variability. Natural speech signals can be grouped into 3 categories depending on the degree of articulation as follow: hypoarticulated, neutral, and hyperarticulated speech. Hypo and hyperarticulation refer to the speech production with respectively a reduction and an increase of the vocal efforts compared to the neutral one [6].

A speaker tends to minimize their articulatory trajectories under hypoarticulation. This relaxed articulatory movement by the reduction of vocal efforts explains many acoustic-phonetic features of hypoarticulated speech. Hypoarticulated speech is characterized by low fundamental frequency (F0), spectral tilting (less energy at higher frequencies), poor harmonics, low amplitude, narrow space between formants, short phoneme duration, less prosody fluctuation, and so on. These acoustic properties often make an acoustic signal phonemically confusable even by a human. Therefore, it can be said that low acoustic contrast and intelligibility under hypoarticulation are the consequence of less vocal efforts. According to the phonetic analysis study reported in [6], hypoarticulation is characterized by high speech rate, fewer syllables, short phoneme duration, more deletion errors, and so on.

On the opposite, hyperarticulation is defined as increasing the articulatory efforts to maximize the clarity of speech. It also shows similar acoustic-phonetic characteristics to those of Lombard speech. Here, Lombard effect refers to the involuntary increment of articulatory efforts in a noisy environment to enhance the audibility of voice [7], [8]. Therefore, the contrary acoustic-phonetic characteristics to hypoarticulation are observed in hyperarticulated speech. The advantage of these properties is to make acoustic speech more intelligible in human communication. However, the speech variability caused by hyperarticulation is not always good for a speech recognition system since the ASR system is primarily modeled by neutral speech data which are easy to be collected. Produced consciously or not, it is well known that the variations of human vocal efforts are associated with the surrounding environment, the communication context, and the conversational intention regard to the listener.

In this work, we adopt several signal processing methods to augment hypo and hyperarticulated speech data as follow: time-scale modification (TSM) for average speaking rate changes, linear signal interpolation/decimation for average pitch modification, the source-filter model of speech production for imitating the articulatory movement of a human, spectral analysis/synthesis for spectral tilt control, SNR-dependent waveform processing for harmonic component enhancement, and so on.

### A. Hypoarticulated Speech Synthesis

Hypoarticulated speech refers to speech produced with the minimization of vocal efforts. It is often observed when a person talks to someone very close. The fluctuation of speech prosody is diminished under hypo-articulation. The average F0 of hypoarticulated speech becomes lower compared to that of neutral speech. And speaking rate becomes fast. In this case, the position of articulatory often moves before reaching the phonetic target. According to the spontaneous speech researches in [3] and [9], it is reported that the articulatory reduction is the main reason for the increase of deletion errors in ASR. And different groups of phonemes are affected differently as speech tempo changes. For example, the average duration of vowels is more reduced from low to fast speech than that of consonants [10].

Among the complex characteristics in hypoarticulated speech, we focus on low F0, articulatory reduction, high speaking rate, and various duration changes. In this work, we adopt linear signal interpolation to realize low pitch frequency and the source-filter model of speech production is applied to mimic the articulatory movement of a human [11], [12]. The traditional variable TSM method is utilized to simplify the complex duration changes depending on different phonetic categories [13]. The increase in the playback speed of speech recoding produces high pitch. Conversely, slowing down the playback speed leads to low F0. Therefore, it is easy to synthesize low pitch signal by exploiting the linear interpolation [14]. Human speech production can be viewed as acoustic filtering operation which shapes spectrum with the airflow from the lungs. Mathematically, a speech signal can be modeled by the combination of sound source and linear filter. And it is actually possible to separate source and filter components from the input. After separating the source and the filter parts, we recreate the articulatory reduction by smoothing the vocal tract filters. In this paper, we deploy the traditional linear predictive coding (LPC) analysis for the source-filter decomposition. And then, line spectral frequencies (LSF's) derived from the LPC's are exploited for smoothing the articulatory trajectories because the stability of LSF's is easily ensured [15]. The variable TSM is an alternative approach for producing more intelligible speech signal via preserving the transient portions of speech [13]. Since speech production involves coordination among articulators with different limitations in movement speed, the different changes in duration between the transient and the steady portions of speech are inevitable. Therefore, it is assumed that the variable TSM method is good for reflecting the characteristics in duration changes. We apply the variable TSM strategy in order to simplify the complex changes in duration depending on phonetic categories.

Fig.1 illustrates the block diagram of the proposed hypoarticulated speech augmentation approach. As shown in Fig.1, we manipulate the source and filter components of speech separately. The output signal is reconstructed by combining the smoothed vocal tract filter and the time-scale modified excitation in the signal synthesis block. Therefore, the LPC and excitation signal synchronization block is necessary. Unlike the original method in [13], transient portions of speech are determined by the cooperation of voice activity detection (VAD) and normalized cross-correlation function (NCCF) in the proposed approach. Simple VAD based on short-time energy threshold is applied to separate speech portions from the input, and

ISBN 978-90-827970-1-5 © EURASIP 2018

NCCF is then exploited to determine steady/transient portions from the speech. Fourier transform based speech synthesis method is deployed for the spectral tilting (less energy at higher frequencies). The spectrum components between 1kHz and 8kHz are linearly suppressed for the spectral tilting. The gain control ($\times 0.5$) in Fig.1 is used to reflect the amplitude reduction under hypoarticulation. In our approach, the output signal is synthesized across the intervals of 10ms. Thus, the temporal trajectory of LSF's needs to be rearranged. We adopt a simple linear prediction technique to estimate LSF's at every 10ms interval. And then, linear smoothing method is applied for the articulatory reduction. By doing so, we can also synchronize the source and filter components in the time domain.

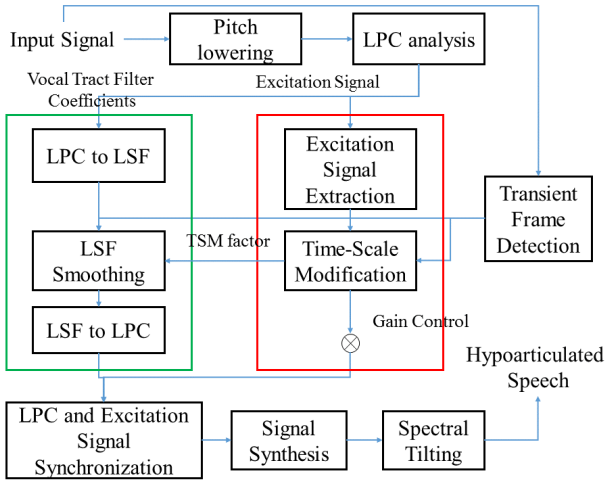


Fig. 1. The block digram of the proposed hypoarticulated speech augmentation approach

### B. *Hyperariculated Speech Synthesis*

In human communication, people often reinforce speech clarity depending on the context of dialogue, the attitude of audiences, the situation of communication, and so on. This increase of vocal efforts leads to high F0, spectral flattening (more energy at higher frequencies), rich harmonics, high amplitude, wide space between formants, long phoneme duration, rich prosody fluctuation, and so on. And speaking rate under hyperarticulation tends to be slow down for the listener's understanding. Therefore, the acoustic contrast for human ear is maximized by the abundant features of hyperarticulated speech. While the variability of speech is hereby increased, the recognition performance under hyperarticulation ironically deteriorates due to the acoustically rich representation of speech. For example, it is reported that insertion errors are increased in recognizing hyperarticulated speech according to the literature in [6]. Among the acoustic-phonetic characteristics of hyper-articulated speech, we spotlight high F0, spectral flattening, slow speaking rate, narrow formant bandwidth, and wide space between formants.

Fig.2 indicates the block diagram of the proposed hyperarticulated speech augmentation approach. We imitate the complex duration phenomena under hyper-articulation by exploiting the traditional variable TSM method [13]. The high F0 and the wide space between formants are realized by the pitch shifting method based on digital signal decimation. We apply the source-filter model of speech production for reflecting the formant bandwidth narrowing and rich harmonics of speech. And speech synthesis method based on spectral analysis is applied to the spectral flattening. In our approach, voiced frames are determined by the cooperation of VAD and normalized auto-correlation function (NACF). After separating speech portions by using the short-time energy based VAD, NACF is utilized to detect voiced frames from the speech. And the traditional SNR-dependent waveform processing method is applied for enhancing the residual harmonics in the voiced speech frame [16]. The formant bandwidth narrowing is implemented by making LSF's closer. The spectrum components between 1kHz and 8kHz are linearly emphasized for the spectral flattening. And the gain control ($\times 1.5$) in Fig.2 is used to reflect the increase in signal amplitude under hyperarticulation.

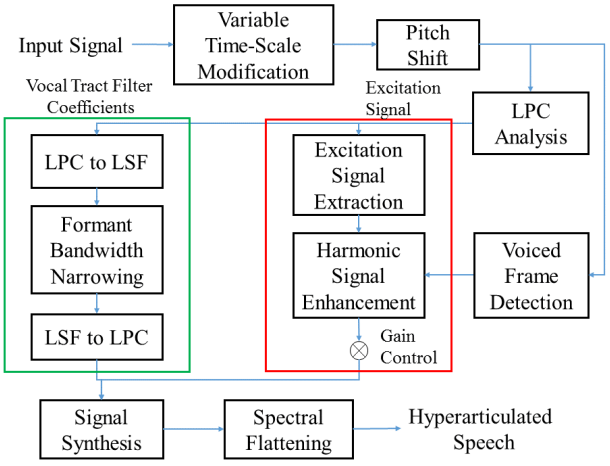


Fig. 2. The block digram of the proposed hyperarticulated speech augmentation approach

## III. Algorithm Implementation

Among the many features in spontaneous speech, we focus on F0, speaking rate, articulatory movement, and formant bandwidth since we think that these attributes are important in the performance of a speech recognition system based on deep neural network (DNN) algorithm.

### A. *Pitch and Speech Rate*

As mentioned before, the low pitch frequency and the fast speaking rate in average are the distinct features of hypoarticulated speech. Conversely, it is obvious that F0 becomes high and speech tempo slows down under hyperarticulation [6]. These acoustic-phonetic attributes can be realized by combining several digital signal processing methods. We are able to easily manipulate pitch frequencies by using signal interpolation and decimation. And it is known that TSM is related with speaking rate change. Although these two methods are simply cascaded, we can achieve a certain target speaking rate by controlling some parameters. For example, 5% signal interpolation leads to 5% pitch lowering but 5% time-domain expansion. If you want to make the speech rate 30% faster with 5% pitch lowing, you can obtain the target speaking rate by the modification factor of 0.65 in TSM. In this work, we adopt the traditional variable TSM method [13] since it can represent the different duration changes depending on phonetic categories even though it's simple manner. The expansion/compression properties of vocalic triangle area

ISBN 978-90-827970-1-5 

reported in [6] is also imitated although it is not perfect. That is, signal interpolation indicates narrowing the area and signal decimation means widening the area.

### B. Articulatory and Formant Bandwidth Reduction

The reduction of articulatory movement is able to explain the acoustic-phonetic characteristics of hypo-articulation. This kind of relaxed speech style often results in the decrease of speech clarity in human speech perception and the increase of deletion errors in ASR. On the other hand, it is reported that the formant bandwidth under Lombard effect becomes narrower than the corresponding one of neutral speech [8]. Therefore, we assume that the increase of vocal efforts under hyperarticulation also results in the sharp formant peaks. In this work, we adopt the source-filter model of speech production to reflect the articulatory reduction of hypoarticulated speech and the sharp formant peak phenomena under hyperarticulation as well. After separating source and filter components from speech, the trajectories of time-varying vocal-tract filters are flattened to mimic the relaxed speaking style. And we make LSF's closer to narrow the formant bandwidth of speech. In this work, total 5 frames including ±2 adjacent frames are smoothed to imitate the minimization of articulation trajectories. We use ±0.01 to make formant peaks sharp under hyperarticulation.

## IV. Experiments

A series of speech recognition test are conducted under several speech styles to assess the feasibility of our approaches in terms of artificial data augmentation: interview, debate, academic presentation, and Lombard speech. Speech data of ~400 hours are prepared for training the acoustic model of a speech recognition system. About 75% portions of the training data are composed of neutral speech such as read speech and the rest are spontaneous speech. The sampling rate of overall speech data is 16kHz. The Kaldi speech recognition toolkit (nnet2) is employed to obtain acoustic models based on popular DNN algorithm [17], [18]. The Gamma-tone filter-bank cepstrum described in [19] and [20] is deployed for the Gaussian mixture model hidden-Markov model (GMM-HMM). And then 40-dimensional log Mel-scale filter-bank energy is used as DNN inputs. The hyper-parameters of our DNN model are as follow: 5 hidden layers with 1,024 nodes, one input layer with 600 (40*(7*2+1)) nodes, and one output layer. The number of output nodes depends on the figure of the GMM-HMM states trained for the forced alignment. That is, the number of output nodes can be altered before and after the data augmentation. However, we control it to be around 8,000 states for a fair evaluation. Linear discriminant analysis without dimensionality reduction is adopted for input normalization. Several speaking styles are considered for the performance evaluation as follow:

1. Test Set 1: interview recordings from Korean broadcast news, 891 files
2. Test Set 2: speech recordings from Korean debate programs, 1,308 files
3. Test Set 3: academic presentation recordings of Korean university students, 1,184 files
4. Test Set 4: Lombard speech from 1.5m distance in noisy environment, 848 files

All the test files are produced by ordinary people and written prompts associated with utterances are not prepared except the set 4. Therefore, it can be said that the test corpus represents each speech styles properly. Although it is difficult to objectively measure the degree of spontaneity, it is assumed that the test sets 1 and 2 contain highly spontaneous speech, whereas the spontaneity of the set 3 is lower compared them. Since the set 4 is collected from 1.5m distance in the presence of ambient noise, the speakers tend to increase their vocal efforts. Therefore, it is presumed that the set 4 stands for Lombard effect.

Table I shows the syllable recognition rates of our baseline speech recognition system. As shown in Table 1, the overall recognition accuracy is low since the quantity of training data is insufficient compared to commercial systems. In addition, the language model is mainly built by using written texts not spontaneous speech style. In addition, the speech recognition system should identify more than unique 498k words.

TABLE I. Syllable Recognition Rates of Baseline Systems

| Training Corpus | Test Set Categories | | | |
|---|---|---|---|---|
| | ***Set 1*** | ***Set 2*** | ***Set 3*** | ***Set 4*** |
| Baseline | 74.97% | 74.34% | 82.48% | 70.96% |

Although we survey many papers, it seems that there is no much knowledge on the pitch change ratios under various speech styles. Therefore, we have to infer the pitch change ratios from the Lombard speech study in the literature [8]. It is also seen that the pitch change ratio under hypoarticulation is lower than that of hyperarticulated speech [6]. The report in [8] indicates that the F0 of Lombard signal is increased by 10.3% in average more than neutral speech. So, we adopt 10% higher F0 for hyperarticulation and 5% lower pitch for hypoarticulation. The vocalic triangle area is also expanded or compressed in proportional to the pitch ratio changes.

According to the literature in [6], the average speech rate (syllable/sec) of hypoarticulated speech becomes 37.3% faster compared to that of neutral speech. Under hyperarticulation, the average speaking rate slows down by 25.5%. In this experiment, we speed up the speech rate by 30% for hypoarticulated speech synthesis. Unlike the slow speaking rate of hyperarticulated speech reported in [6], we speed up the target speaking rate by 20% for hyperarticulation since it seems that fast speech rate more affects a speech recognition system than the slow one [21], [22]. In addition, it is observed from the test corpus (Test Set 1-3) that the speaking rate under hyperarticulation is still high because Koreans are fast speakers.

In this experiment, the half of the training data are synthesized to represent hypoarticulation and the other half are used for mimicking hyperarticulated speech. Table II indicates the syllable recognition rates of the proposed data augmentation approaches. As shown in Table II, we achieve the 2.76% absolute improvement in speech recognition by augmenting hypo and hyperarticulated speech in the case of the test set 2. According to the study in [3], it is shown that the portions of hypo and hyperarticulation in total

ISBN 978-90-827970-1-5 © EURASIP 2018

spontaneous speech is relatively small (See Fig.1 in the reference [3]). Therefore, we think that the 2.76% absolute improvement is meaningful even though its error reduction rate is not dramatic.

TABLE II. SYLLABLE RECOGNITION RATES AFTER DATA AUGMENTATION

| Training Corpus | Test Set Categories | | | |
|---|---|---|---|---|
| | *Set 1* | *Set 2* | *Set 3* | *Set 4* |
| AUG. | 77.16% | 77.10% | 83.59% | 71.59% |
| ERR | 9% | 11% | 6% | 2% |

Unlike the spontaneous speech styles (Test Set 1-3), the average phoneme duration tends to be extended in the case of Lombard speech. Therefore, the performance gain for the test set 4 is not relevant. However, we obtain 1.81% absolute performance gain by increasing hyperarticulated speech data only. In that test, we slow down the target speaking rate by 10%. This experiment indicates that the more matched data, the better recognition result.

## V. CONCLUSIONS

Improving ASR performance in spontaneous speech is known as a very difficult task. Despite much research effort dedicated, only a small reward is often achieved. And there are many challenges in automatically recognizing spontaneous speech using a machine. In this paper, we focus on the variability of speech depending on the different degrees of articulation (e.g. hypo and hyperarticulation phenomena). We assume that several distinct characteristics affect the performance of a speech recognition system: pitch, harmonics, formant, duration, spectral tilt, speech rate, and articulatory movement. In this work, we develop speech augment approaches producing hypo and hyperarticulated speech from the neutral speech corpus. Since our approaches are based on simple signal processing methods we are able to manipulate speech parameters directly. And the feasibility of our approaches is investigated in terms of ASR. A series of speech recognition tests across several speech styles are conducted to evaluate the proposed approaches. The result shows that we achieve meaningful performance gain in the natural speech recognition task. This indicates that the acoustic-phonetic features reported in the past studies related to of hypo and hyperarticulated speech are useful for dealing with the variability of speech in spontaneous speech recognition.

## ACKNOWLEDGMENT

This work was supported by the ICT R&D program of MSIT/IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning].

## REFERENCES

[1] S. Furui, “Recent advances in spontaneous speech recognition and understanding,” in Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, pp. 1-6, April 2003.

[2] S. Furui, Toward spontaneous speech recognition and understanding. In: Pattern Recognition in Speech and Language Processing. W. Chou, B.-H. Juang, CRC Press, New York, pp. 191-227, 2003.

[3] M. Nakamura, S. Furui, K. Iwano, “Acoustic and Linguistic Charaterization of Spontaneous Speech,” ITRW Workshop on Speech Recognition & Intrinsic Speech Variation (SRIV 2006), pp. 3-8, May 2006.

[4] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, Josef Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, “Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives,” IEEE Trans. on Speech and Audio Processing, vol. 12, no. 4, pp.420-435. July 2004.

[5] S. Furui, “Introduction to the Special Issue on Spontaneous Speech Processing,” IEEE Trans. on Speech and Audio Processing, vol. 12, no. 4, pp. 349-350, July 2004.

[6] B. Picart, T. Drugman, T. Dutoit, “Analysis and HMM-based syhthesis of hypo and hyperarticulated speech,” Computer Speech & Language, vol. 28, issue 2, pp. 687-707, March 2014.

[7] H. Lane and B. Tranel, “The Lombard Signal and the Role of Hearing in Speech,” Journal of Speech, Language, and Hearing Research, vol. 14, pp. 677-709, Dec. 1971.

[8] D. Huang, R. Susanto, E. P. Ong, “Lombard Effect Mimicking,” in Proc. International Conference on Spoken Language Processing (INTERSPEECH), pp. 258-263, Sep. 2010.

[9] S. Furui, “Selected Topics from 40 Years of Research on Speech and Speaker Recognition,” in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1-8, Sep. 2009.

[10] H. Kuwabara, “Perceptual and Acoustic Properties of Phonemes in Continuous Speech for Different Speaking Rate,” in Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1003-1006, Sept. 1997

[11] G. Fant, Acoustic Theory of Speech Production, The Hague, Mouton Netherlands, 1960.

[12] K. N. Stevens, Acoustic Phonetics, The MIT Press, London England, 1998.

[13] S. Lee, H Kim, H Kim, “Variable time-scale modification of speech using transient information,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1319-1322, April 1997.

[14] R. W. Schafer and L. R. Rabiner, “A Digital Signal Processing Approach to Interpolation,” Proc. IEEE, vol. 61, pp. 692-702, July 1973.

[15] K.K. Paliwal, “A study of line spectrum pair frequency for speech recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New York, pp. 485-488, April 1988.

[16] D. Macho and Y. M. Cheng, “SNR-Dependent Waveform Processing for Improving the Robustness of ASR front-end,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 305-308, May 2001.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, “The Kaldi speech recognition toolkit,” IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Hawaii, 2011.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A, Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[19] S. Lee, B. Kang, H. Chung, and Y. Lee, “Intra- and inter-frame features for automatic speech recognition,” ETRI Journal, vol. 36, no. 3, pp. 514-517, June 2014.

[20] S. Lee, B. Kang, H. Chung, and Y. Lee, “A useful feature-engineering approach for an LVCSR system based on CD-DNN-HMM algorithm,” The 2015 European Signal Processing Conference (EUSIPCO 2015), pp. 1436-1440, Sept. 2015.

[21] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi. C. Wellekens, “Automatic speech recognition and speech variability: A review,” Journal of Speech Communication, vol. 49, no. 10-11, pp. 763-786, Oct.-Nov. 2007.

[22] N. Morgan, E. Fosler, and N. Mirghafori, “Speech recognition using on-line estimation of speaking rate,” European Conference on Speech Communication and Technology (EUROSPEECH), pp. 2079-2082, Sept. 1997.