Improving Automatic Forced Alignment for Dysarthric Speech Transcription

Yu Ting Yeung[2], *Ka Ho Wong*[1], *Helen Meng*[1,2]

[1]Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management
[2]Stanley Ho Big Data Decision Analytics Research Centre
The Chinese University of Hong Kong, Hong Kong SAR, China
{ytyeung,khwong,hmmeng}@se.cuhk.edu.hk

Abstract

Dysarthria is a motor speech disorder due to neurologic deficits. The impaired movement of muscles for speech production leads to disordered speech where utterances have prolonged pause intervals, slow speaking rates, poor articulation of phonemes, syllable deletions, etc. These present challenges towards the use of speech technologies for automatic processing of dysarthric speech data. In order to address these challenges, this work begins by addressing the performance degradation faced in forced alignment. We perform initial alignments to locate long pauses in dysarthric speech and make use of the pause intervals as anchor points. We apply speech recognition for word lattice outputs for recovering the time-stamps of the words in disordered or incomplete pronunciations. By verifying the initial alignments with word lattices, we obtain the reliably aligned segments. These segments provide constraints for new alignment grammars, that can improve alignment and transcription quality. We have applied the proposed strategy to the TORGO corpus and obtained improved alignments for most dysarthric speech data, while maintaining good alignments for non-dysarthric speech data.

Index Terms: automatic forced alignment, speech recognition, dysarthric speech, word lattices

1. Introduction

Dysarthria is a motor speech disorder due to problems in the nervous system [1]. The disorder is usually a consequence of diseases such as stroke, cerebral palsy and amyotrophic lateral sclerosis (ALS). These diseases also cause impaired movement of the limbs, prohibiting the use of sign languages, computer input by keyboard or pointing devices. The muscular impairment also affects speech production which leads to disordered speech. The impairment presents great difficulty to the subjects in daily communication.

The recent advances of speech technologies bring a new hope to dysarthric subjects on improving their quality of life, such as using their remaining speech abilities to perform spoken control of daily tasks [2–4], speech synthesis and voice conversion for more lively and animated expression [5, 6], and computer programs for speech rehabilitation [7]. Today's speech technologies are mostly data-driven and rely on speech corpora. For dysarthric speech, several English dysarthric speech corpora are publicly available, such as Nemours [8], the Universal Access-Speech (UA-Speech) [9], and the TORGO corpus [10].

Forced alignment automatically generates time-stamps of linguistic units (e.g. words, syllables, or phonemes) according to the transcriptions of the recordings. The characteristics of dysarthric speech pose challenges for automatic forced alignment. The automatic forced alignment algorithm should be robust to the prolonged pause intervals and slow speaking rates. The algorithm should also be robust to the poor articulation of phonemes, insertion and deletion of syllables and phones. Moreover, the subjects may delete a word or restart an utterance. The spoken content in the recordings may be different from transcriptions.

Manual alignment of dysarthric speech data is a difficult task due to poor speech intelligibility. Automatic forced alignment of dysarthric speech is thus preferable. We may apply a well-trained acoustic model (AM) from normal (i.e., non-dysarthric) speech data for automatic forced alignment of dysarthric speech. Noticing the risk of speech data mismatch, we must verify the quality of the alignments. Dysarthric speech usually contain long pauses. We first perform initial alignment to locate all the long pauses in the dysarthric utterances. Previous works [11–13] successfully applied large vocabulary continuous speech recognition (LVCSR) in aligning very long transcriptions. We adopt the speech recognition approach for the flexibility of inserting and deleting words in the alignments. The word lattices from LVCSR system also provide us with larger search spaces to locate the reliably aligned segments. The reliably aligned segments provide the constraints for new alignment grammars that aim at improving the alignments.

This paper is organized as follows. In Section 2, we briefly describe our speech data. We also discuss the challenges of aligning dysarthric speech. In Section 3, we present our method to obtain and update word alignments on dysarthric speech. In Section 4, we discuss the alignment results on both dysarthric and non-dysarthric speech. Finally in Section 5, we present our conclusions and future directions.

2. Dysarthric speech

2.1. The corpus

We use the TORGO (LDC2012S02) corpus [10] for dysarthric speech data. The corpus includes 8 dysarthric subjects (5 males and 3 females) and 7 non-dysarthric speakers (4 male and 3 females). The corpus includes action tasks for articulation movements, speaking tasks of repeating patterns, words, sentences and picture descriptions. We only use speech data from single-word and sentence speaking tasks as transcriptions are available. The data set consist of about 4,900 non-dysarthric and 2,400 dysarthric speech utterances, with at least 100 utterances

This project is partially sponsored by a grant from the HKSAR Government General Research Fund (ref. no. 415513).

Appears in INTERSPEECH 2015 pp. 2991-2995
Updated version with typos corrected

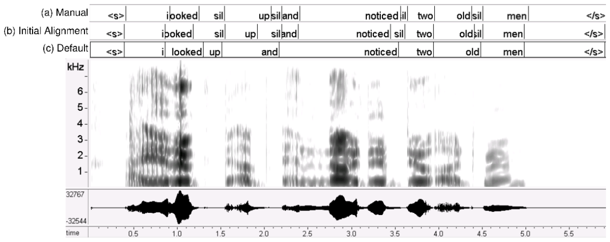

Figure 1: A comparison of different alignments on dysarthric speech (M01/Session2_3/0094.wav). The waveform and spectrogram of the utterance are included. (a) Manual alignment. (b) Pause-aware initial alignment. (c) Conventional automatic forced alignment without handling the pauses. A pause is represented as “sil”. <s>and </s> represent the sentence start and end of the utterance respectively.

Table 1: The statistics of pause interval length in dysarthric and non-dysarthric speech.

	Non-dysarthric	Dysarthric
Mean (ms)	74	383
25th percentile (ms)	30	80
Median (ms)	50	200
75th percentile (ms)	80	540

for each speaker. The speech data are recorded with head-mounted microphones. The TORGO corpus also includes Frenchay Dysarthria Assessment (FDA) [14] results of the subjects, which provide references to the severity of dysarthria.

2.2. The challenges of aligning dysarthric speech

2.2.1. Prolonged pause intervals

The existence of long pauses is characteristic of dysarthric speech. Table 1 shows the statistics of pause intervals of the TORGO corpus from the pause-aware initial alignment stage described in Section 3.1. For non-dysarthric speech, the pause intervals are usually short. The length is consistent among different speakers. These pauses are commonly modeled as “short pause” (“sp tee-model” in HTK toolkit [15]) in acoustic modeling. The pause intervals of dysarthric utterances are significantly longer. The length of a pause is comparable to that of a syllable. We need to take extra care on these pause intervals during alignment. Figure 1 shows the waveform, spectrogram and alignments from different alignment methods of a dysarthric utterance. The speaker pauses nearly at every word, and even between two syllables within the word “noticed”. The longest pause lasts for nearly 300 ms, leading to a slow speaking rate of about 2 syllables per second. The pause-aware initial alignment (b) generally agrees with the manual alignment (a). When we apply the conventional forced alignment setting (“short pause” model and transcriptions without pause markings) as in (c), alignment errors are observed (between 1.3-2.4 s, for the words “up” and “and”).

The pause intervals are as important as words in aligning dysarthric speech. The pause intervals in dysarthric speech can be modeled well by the pause model derived from the acoustic model of normal speech, due to similar acoustic properties. Once a pause interval is located, it can serve as an anchor point in the alignment. The start-time of a pause is the same as the end-time of the previous word. The end-time of a pause is the same as the start-time of the next word.

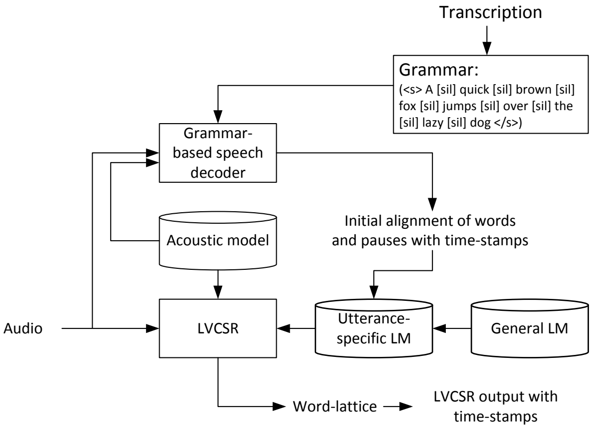
Transcription
Grammar:
(<s> A [sil] quick [sil] brown [sil] fox [sil] jumps [sil] over [sil] the [sil] lazy [sil] dog </s>)
Grammar-based speech decoder
Initial alignment of words and pauses with time-stamps
Acoustic model
Audio
LVCSR
Utterance-specific LM
General LM
Word-lattice
LVCSR output with time-stamps

Figure 2: The flow diagram of the grammar-based speech decoder for generating initial alignments. The flow diagram also includes the LVCSR system for generating word lattices and speech recognition outputs. An optional pause is represented by [sil] in the grammar.

2.2.2. Pronunciation deviations

Pronunciation deviations are common in dysarthric subjects. The physical limitations of the subjects lead to poor articulation of phonemes, insertion and deletion of syllables and phones. The observed pronunciation deviation patterns of the TORGO corpus are described in [4]. We decide to perform alignment at word-level as the linguistic boundaries are better defined by the pause intervals. The contextual information from language models in LVCSR also helps to compensate for the acoustic mismatch due to pronunciation deviations. We may perform phone-level alignment at each word after segmenting a sentence into words according to the word-level alignment.

3. Alignment of dysarthric speech

3.1. Stage 1: Pause-aware initial alignment

The first stage is to locate the long pauses in dysarthric utterances. We recognize each utterance with a grammar-based speech recognizer based on the HTK toolkit [15]. The grammar is built upon the original word-level transcription, but optional pauses are allowed to be inserted between words. We refer this stage to as initial alignment. Figure 2 shows the corresponding flow diagram and an example of the decoding grammar. We prepare a speaker-independent acoustic model (AM) for the initial alignment task. The AM is trained with training data of the TIMIT (LDC93S1) corpus [16] and represents the phonetic characteristics of normal speech from a wide variety of speakers. The monophone AM consists of 128 Gaussian components at each acoustic state.

3.2. Stage 2: LVCSR output

We adopt the approach of using LVCSR outputs to verify the initial alignments [11–13]. Figure 2 also shows the flow diagram of the LVCSR system. We use the same 128-state TIMIT monophone AM. We do not use a triphone AM, as speech recognition performance is similar on dysarthric speech [4].

General word-based bigram and trigram language models (LMs) are prepared from the entire original transcriptions of the TORGO corpus. For each utterance, we further interpolate the general bigram and trigram LMs with the text of the ini-

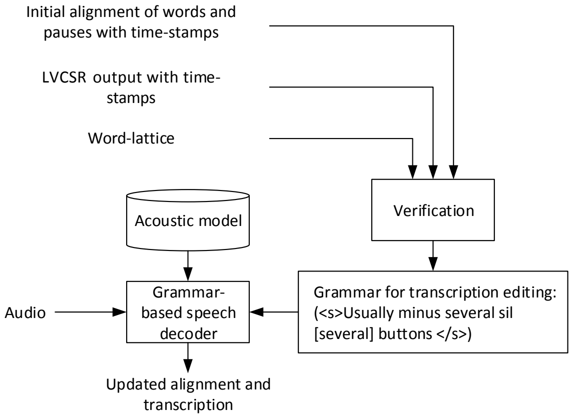
Initial alignment of words and pauses with time-stamps
LVCSR output with time-stamps
Word-lattice
Verification
Acoustic model
Audio
Grammar-based speech decoder
Grammar for transcription editing: (<s>Usually minus several sil [several] buttons </s>)
Updated alignment and transcription

Figure 3: The flow diagram of the alignment verification stage and the grammar-based speech decoder for updating the alignments. An example of update grammar is given, where an optional word is braced by brackets [].

tial alignment for utterance-specific LMs. We include the pause markings into the utterance-specific LMs to model the pause intervals in the utterance. The interpolation weight is set to 0.9, to bias toward the text of the initial alignment.

A two-pass LVCSR is developed with the HTK toolkit [15]. For each utterance, word lattice is generated with the utterance-specific bigram LM. The word lattice is then re-scored by the utterance-specific trigram LM for 1-Best LVCSR output. The 1-Best output accompanies with time-stamps of the recognized text. When the text of the 1-Best output (pauses ignored) is the same as the original transcription, the original transcription is considered as reliable and the 1-Best output is the final alignment of the utterance.

3.3. Stage 3: Alignment verification and update

The 1-Best LVCSR output may be different from the original transcription. We have to identify whether the difference is due to speech recognition errors or imperfect original transcription. Speech recognition errors of dysarthric speech are commonly caused by pronunciation deviations of the subjects. Subjects may also pause between syllables of a word. A longer word may be recognized as several shorter words. We thus process the 1-Best output by grouping the recognized text into segments. The time-stamps of the segments are aligned with the time-stamps of the words and the pauses in the initial alignment. We compute the Levenshtein distance between the segments and the text (including the pauses) of the initial alignment for the mismatch patterns, i.e., insertion, deletion and substitution. We weight equally on all types of mismatch patterns. We also add an additional constraint for pattern matching. The matched patterns should also match with either start-time or end-time.

A word lattice contains alternative paths in addition to the 1-Best output. We hope that a substituted word can be recovered from the word lattice. When we discover an arc of the substituted word with the same end-time as the initial alignment, we consider the word as reliable. We match only the end-time to allow potential word insertion in the updated alignment. A grammar for updating the alignment is generated according to the rules described in the following. The rules make use of the mismatched patterns, word reliability, and the time-stamps of the substituted and recognized words. We allow a maximum of 200 ms tolerance for time-stamp matching between the initial alignments and the 1-Best outputs. The rules are:

1. If a substituted word is considered as reliable, the word is always recovered in the updated alignment.
2. If the substituted word is reliable, we replace the recognized word with the substituted word in the updated alignment under the following conditions: (1) the start-time *or* the end-time between the substituted and the recognized words are matched, or (2) the start-time *and* the end-time are matched with larger time tolerance.
3. Following Rule 2, when the substituted word is unreliable, the word may be replaced by the recognized word.
4. If the recognized word ends after the substituted word, or vice versa, the updated alignment may include both words. The word order follows the end-time of the words. The recognized word is always optional in the updated alignment.
5. An inserted word in the 1-Best output is optional in the updated alignment.
6. If a word in the initial alignment is deleted in the 1-Best output, the word is optional in the updated alignment.

An extension is especially useful for handling the pronunciation deviations of dysarthric speech. We may abolish Rule 3 and extend Rule 2 to replace the recognized word with the substituted word as long as their time-stamps are matched. Finally, we re-align the utterances with a grammar-based speech recognizer according to the derived grammar.

4. Result analysis

4.1. Agreement of LVCSR outputs

We align the dysarthric speech data with the settings described in Section 3. We define the agreement rate which is equivalent to sentence correctness (1- sentence error rate) of the 1-Best LVCSR outputs (pauses ignored) with the original transcriptions as the reference. We expect that the agreement rate should be reasonably high when aligning speech data in a carefully designed corpus. We do not expect a 100% agreement rate. An occasional mismatch between the spoken content and the transcription can occur in a well-designed corpus.

A hyper-parameter of LVCSR speech decoder is the language model (LM) weight. The LM weight controls the contribution of acoustic similarity according to acoustic model and contextual information from the utterance-specific LMs. We have considered two LM weights. A lighter LM weight $p = 15$ is within the common range of LVCSR systems, and a heavier LM weight $p = 30$. Table 2 shows the agreement rates of individual speakers.

For non-dysarthric speech data, the 1-Best LVCSR outputs generally agree with the original transcriptions regardless the LM weights. The medians of the agreement rates are 91.89% for $p = 15$ and 95.09% for $p = 30$. The over 90% agreement rate suggests that the original transcriptions are generally in good quality. The high agreement rates also suggest that the acoustic mismatch between the TORGO and the TIMIT corpora is negligible for non-dysarthric speech. The agreement rates are also consistent among different non-dysarthric speakers.

The LVCSR agreement rates drop substantially in dysarthric speech data. The medians of the agreement rates are 61.7% and 85.1% for $p = 15$ and 30 respectively. The situation is expected due to the distorted pronunciations. For lighter LM weight $p = 15$, the agreement rates are coarsely related to

Table 2: 1-Best LVCSR output agreement rates of individual speakers. The severity of the subjects are described in [4].

Dysarthric subjects				Non-dysarthric subjects		
Speakers	Severity	Agreement rate (%)		Speakers	Agreement rate (%)	
		p=15	p=30		p=15	p=30
F01	Severe	63.16	87.72	FC01	91.89	95.27
F03	Moderate	80.91	88.18	FC02	96.27	96.69
F04	Mild	84.68	89.52	FC03	92.17	95.09
M01	Severe	53.53	79.35	MC01	93.05	95.96
M02	Severe	60.26	82.56	MC02	81.31	90.99
M03	Mild	90.12	91.60	MC03	88.54	89.17
M04	Severe	54.74	74.82	MC04	91.36	92.48
M05	Moderate-to-severe	26.55	79.65			

Table 3: Agreement rates between the updated alignments and the original transcriptions of individual speakers.

Dysarthric subjects			Non-dysarthric speakers		
Speakers	Agreement rate (%)		Speakers	Agreement rate (%)	
	(a)	(b)		(a)	(b)
F01	92.11	94.74	FC01	98.65	98.65
F03	94.90	94.90	FC02	99.27	99.27
F04	94.35	95.16	FC03	97.29	97.49
M01	86.96	90.49	MC01	98.40	98.50
M02	92.01	93.56	MC02	94.14	94.14
M03	97.28	97.52	MC03	92.43	92.43
M04	85.71	87.91	MC04	97.28	97.28
M05	85.84	90.27			

the severity of the subjects. The mild subjects achieve higher agreement rates, although the rates are still lower than those of non-dysarthric speakers. For the severe subjects, their agreement rates are generally lower. M05 gets the lowest agreement rate although the severity level is not the worst. This is probably due to serious pronunciation deviations of the subject. A note in the TORGO corpus states that the speech of M05 is intelligible when the subject speaks slowly.

A heavier LM weight $p = 30$ increases the agreement rates of the dysarthric subjects. The agreement rates become more consistent among different dysarthric subjects and no longer reflect their severity. The severe subjects tend to achieve larger improvement on the agreement rates with the heavier LM weight. The agreement rate of M05 is no longer the lowest. The pronunciation deviations of the subjects are compensated by the strong contextual information from the utterance-specific LMs.

4.2. Alignment agreement after verification and update

We proceed to the verification stage with the LVCSR outputs based on LM weight $p = 30$ due to the higher and more consistent LVCSR agreement rates. We compare two different settings for the alignment verification process: (a) the basic setting described in Section 3.3, and (b) the extension in which the recognized words are replaced by the substituted words when their time-stamps are matched. Table 3 shows the agreement rates between the updated alignments and the original transcriptions. The agreement rates of non-dysarthric speakers are close to 100%. The results from different settings are similar. For dysarthric subjects, the medians of the agreement rates increase to 92.4% and 94.2% for setting (a) and (b) respectively.

Figure 4 shows the example outputs of three utterances. The first utterance is from a non-dysarthric speaker. The speaker paused shortly and restarted the sentence. The 1-Best LVCSR output detects the restart at the word "several" although the word "buttons" is wrongly recognized. The verification step recovers the word "buttons". The grammar also hypothesizes the inserted word "several". The updated alignment now matches with the ground truth. In the second utterance, we revisit the example of Figure 1. The utterance is from a subject with severe dysarthria. Pronunciation derivation is observed at the word "looked". The word is recognized as "two". The first half of the 1-Best LVCSR output is erroneous, but the reliable segments of the initial alignment are still identified from the word lattice. After the alignment verification, most part of the new alignment matches with the ground truth, but the word "looked" is not recovered under setting (a). The word can be recovered under setting (b), due to the matched time-stamps between "looked" and "two". The third utterance is spoken by another severe subject. Although there are some recognition errors in the 1-Best LVCSR output, the word lattice reveals that the initial alignment is reliable. The initial alignment is accepted as the final alignment output. As shown in the examples, a pause is a reliable feature in dysarthric speech. The pauses are always included in the grammar to update the alignments.

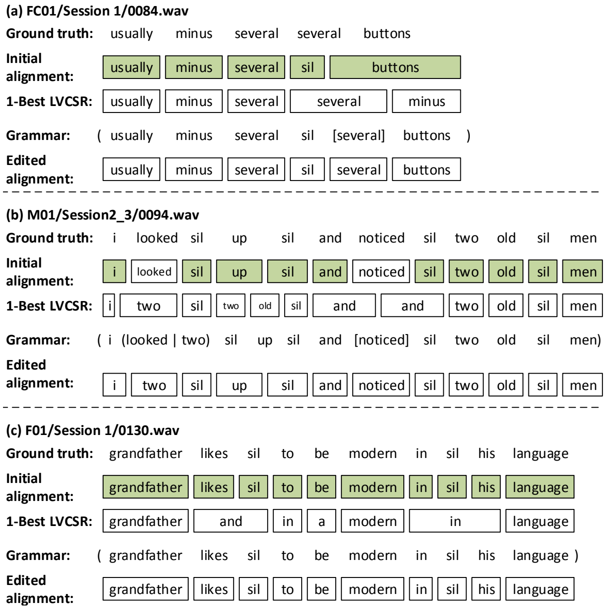
(a) FC01/Session 1/0084.wav
Ground truth: usually minus several several buttons
Initial alignment: usually minus several sil buttons
1-Best LVCSR: usually minus several several minus
Grammar: (usually minus several sil [several] buttons)
Edited alignment: usually minus several sil several buttons
(b) M01/Session2_3/0094.wav
Ground truth: i looked sil up sil and noticed sil two old sil men
Initial alignment: i looked sil up sil and noticed sil two old sil men
1-Best LVCSR: i two sil two old sil and and two old sil men
Grammar: (i (looked | two) sil up sil and [noticed] sil two old sil men)
Edited alignment: i two sil up sil and noticed sil two old sil men
(c) F01/Session 1/0130.wav
Ground truth: grandfather likes sil to be modern in sil his language
Initial alignment: grandfather likes sil to be modern in sil his language
1-Best LVCSR: grandfather and in a modern in language
Grammar: (grandfather likes sil to be modern in sil his language)
Edited alignment: grandfather likes sil to be modern in sil his language

Figure 4: Examples of ground truths, initial alignments, LVCSR outputs and updated alignments. The length and the location of the boxes illustrate the time-alignment of the words in the utterances. In the initial alignments, reliable words found in word lattices are shaded. In the grammars, an optional word is braced by brackets [], and a choice is represented by (|).

5. Conclusions

We have developed a strategy for automatic forced alignment on dysarthric speech data with an acoustic model trained with normal speech data. We have applied the strategy to align the TORGO corpus and verified a set of reliable alignments of dysarthric speech data. These alignments[1] should be of good quality to support research in dysarthric speech. We continue to improve the alignment strategy to process the remaining TORGO speech data and other dysarthric speech corpora.

[1] Alignments available: https://github.com/ytyeung/IS2015alignment

6. References

[1] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 3rd ed. Elsevier Health Sciences, 2013.

[2] M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green, "Automatic speech recognition and training for severely dysarthric users of assistive technology: The stardust project," *Clinical Linguistics & Phonetics*, vol. 20, no. 2-3, pp. 149–156, 2006.

[3] M. Hasegawa-Johnson, J. Gunderson, A. Penman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *Acoustics, Speech and Signal Processing (ICASSP), 2006. IEEE International Conference on*, 2006, pp. 1060–1063.

[4] K. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011. IEEE International Conference on*, 2011, pp. 4924–4927.

[5] A. Kain, X. Niu, J.-P. Hosom, Q. Miao, and J. P. v. Santen, "Formant re-synthesis of dysarthric speech," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 25–30.

[6] J.-P. Hosom, A. Kain, T. Mishra, J. van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2003. IEEE International Conference on*, 2003, pp. 924–927.

[7] A. Hatzis, P. Green, J. Carmichael, S. Cunningham, R. Palmer, M. Parker, , and P. O'Neill, "An integrated toolkit deploying speech technology for computer based speech training with application to dysarthric speakers," in *EUROSPEECH-2003*, 2003, pp. 2213–2216.

[8] X. Menendez-Pidal, J. Polikoff, S. Peters, J. Leonzio, and H. Bunnell, "The Nemours database of dysarthric speech," in *The Fourth International Conference on Spoken Language Processing (ICSLP)*, vol. 3, Oct 1996, pp. 1962–1965 vol.3.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH 2008*, 2008, pp. 1741–1744.

[10] F. Rudzicz, A. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[11] P. J. Moreno, C. F. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *The Sixth International Conference on Spoken Language Processing (ICSLP)*, 1998, pp. 2711–2714.

[12] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings." in *INTERSPEECH 2006*, 2006, pp. 1606–1609.

[13] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Acoustics, Speech and Signal Processing (ICASSP), 2009. IEEE International Conference on*, 2009, pp. 4869–4872.

[14] P. M. Enderby, *Frenchay dysarthria assessment*. San Diego, California: College-Hill Press, 1983.

[15] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University, 1995.

[16] Garofolo, John, et al., "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Web download. Philadelphia: Linguistic Data Consortium, 1993.