INTELLIGIBILITY OF MODIFICATIONS TO DYSARTHRIC SPEECH

John-Paul Hosom[1], *Alexander B. Kain*[1], *Taniya Mishra*[1], *Jan P. H. van Santen*[1], *Melanie Fried-Oken*[2], *and Janice Staehely*[2]

[1]Center for Spoken Language Understanding, OGI School of Science & Engineering
[2]Assistive Technology Program, Child Development and Rehabilitation Center
Oregon Health & Science University, Portland, Oregon, USA
{hosom,kain,mishra,vansanten}@ogi.edu, {friedm,staehely}@ohsu.edu

ABSTRACT

Dysarthria is a motor speech impairment affecting millions of people. Dysarthric speech can be far less intelligible than that of non-dysarthric speakers, causing significant communication difficulties. The goal of this work is to understand the effect that certain modifications have on the intelligibility of dysarthric speech. These modifications are designed to identify aspects of the speech signal or signal processing that may be especially relevant to the effectiveness of a system that transforms dysarthric speech to improve its intelligibility. A result of this study is that dysarthric speech can, in the best case, be modified only at the short-term spectral level to improve intelligibility from 68% to 87%. A baseline transformation system using standard technology, however, does not show improvement in intelligibility. Prosody also has a significant ($p < 0.05$) effect on intelligibility.

1. INTRODUCTION

A 1992 survey reported that there are at least 2.5 million adult Americans with significant disability due to chronic neurologic impairment, including Parkinson's, Multiple Sclerosis, and stroke. A large percentage of these people present with *Dysarthria*, a motor speech impairment due to weakness or poor coordination of the muscles used in speech production. People with dysarthria can communicate via typing or speech; a person who relies on augmentative or alternative communication devices to communicate typically types words at a rate 150 to 300 times slower than average [1]. The reduced intelligibility of dysarthric speech can make verbal communication laborious, time-consuming, and frustrating.

Because dysarthric speech is generally hard to understand, specialized speech transformation techniques [2] have been developed to make it more intelligible, thereby enabling individuals with dysarthria to communicate more effectively. However, these techniques are not obviously different from spectral filters and amplifiers that enhance certain parts of the spectrum, and are hence unlikely to be of help in cases where phoneme production is seriously impaired (e.g., pronouncing [t] as [n]) or where the overall pitch and loudness pattern is disrupted. Our group is working on a new type of intelligibility-enhancing system that addresses these more serious impairments, making use of a combination of speech recognition, speech transformation, and synthesis methods (Figure 1). Our baseline speech transformation system uses standard techniques from voice transformation that map the spectral characteristics of one speaker to be more similar to those of another speaker. (We use the term "speech transformation" here for referring to a system that transforms speech for the goal of improved intelligibility, and the term "voice transformation" for referring to a system that has the goal of modifying speaker identity.)

A key feature of the proposed architecture is the separation between the roles of prosodic and spectral information. Existing voice transformation technology typically operates only on the speech spectrum and leaves detailed prosodic factors (other than global fundamental frequency (F_0) and energy values) untouched; hence, it has not been clear to what degree research efforts should focus on (a) the *Prosody Extractor* and the integration of this information into the *Speech Transformer* or (b) improvement of the existing capabilities of the *Speech Transformer*, possibly based on integration of phoneme- and word-level recognition or more robust feature spaces. Previous work [3] has shown that at least the F_0 component of prosody can play a significant role in intelligibility of dysarthric speech, but there have not been studies that compare all aspects of prosody simultaneously with spectral characteristics.

The goal of this work is to understand the effect that certain modifications have on the intelligibility of dysarthric speech. These modifications, described in more detail in Section 4, have been designed to identify aspects of the speech signal or signal processing that may be especially relevant to the effectiveness of a speech transformation system. The modifications address prosody, spectral content, regions of the signal containing formants, and general effects of signal processing. The research directions of a future speech transformation system can then be guided by the results of this study.

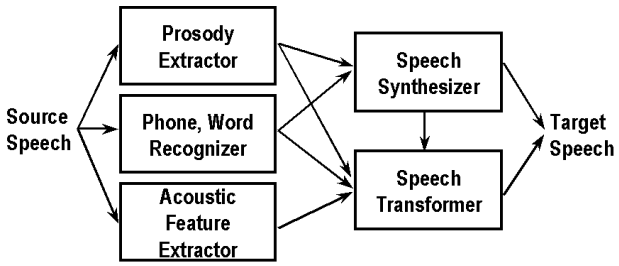

Fig. 1. System architecture for improving intelligibility of dysarthric speech.

This research was conducted with support from NSF Grant 0117911 "Making Dysarthric Speech Intelligible".

2. SPEECH DATA

The speech data used for training and evaluation consisted of one dysarthric speaker and one non-dysarthric ("normal") speaker from the Nemours database of dysarthric speech [4].

The Nemours database contains a number of speakers reading sentences and paragraphs. The sentences are of the form "The X is Ying the Z," where the words X, Y, and Z are syntactically correct but semantically vacuous. One such sentence is "The dive is singing the phase." The words have been chosen according to criteria specified in Kent *et al.* [5] in order to facilitate an understanding of what types of speech are less intelligible for a given speaker. There are 74 sentences per speaker; the first 37 sentences contain the same words as the last 37 sentences, but with the order of the two nouns X and Z reversed. For this study, we used only the sentences from speaker LL as the *source speaker*. A non-dysarthric speaker, JP, uttered the same sentences; this speaker was the *target speaker* for our speech-transformation experiments and the second speaker for the "hybrid" stimuli described in Section 4.

We have manually labeled the sentences spoken by LL and JP with phoneme identity and time alignments in order to ensure quality phonetic segmentations of each utterance.

3. SPEECH TRANSFORMATION SYSTEM

3.1. Analysis and Synthesis

Our baseline speech transformation system works on the frame level, without any knowledge of context. During transformation, each frame is first analyzed to extract its F_0, energy, and spectral features. These features are then transformed using a model that has been trained to learn the relationship between dysarthric and normal speech, and a new speech signal is reconstructed during synthesis. Details of the analysis and synthesis steps follow, while transformation is discussed in Section 3.3.

In the analysis step, we first divide the speech, sampled at 16 kHz, into 30 ms overlapping frames, at a rate of 5 ms. For each frame, we calculate energy and 24^{th} order real cepstral coefficients via

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \\ \hat{X}(k) &= \log|X(\mathrm{B}(k))| \\ c[n] &= \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}(k) e^{j\frac{2\pi}{N}kn} \end{aligned}$$

where $x[n]$ represents the windowed speech frame, $X[k]$ its N-point discrete Fourier transform, B the bark-scale warping function, and $c[n]$ the discrete real cepstrum. F_0 and voicing information are calculated using WaveSurfer [6].

During synthesis, the low-quefrencies of the liftered cepstrum are used to estimate the original spectrum

$$\begin{aligned} \hat{X}(k) &= \sum_{n=0}^{24} c[n] e^{-j\frac{2\pi}{N}kn} \\ X(\mathrm{B}(k)) &= e^{\hat{X}(k)} \end{aligned}$$

Next, we calculate a high-order LPC filter for every frame by applying the Levinson recursion to the biased autocorrelations. The LPC filters are then excited by a train of impulses in voiced regions, and white noise in unvoiced regions, according to the desired F_0 and binary voicing features. Finally, the energies of the resulting frames are adjusted to match the specified energies.

Note that this method does not reconstruct the original speech signal perfectly. The resynthesized speech is different from the original in that its magnitude spectrum has been smoothed and its phase spectrum has been replaced by either minimum or random phase (according to the binary voicing feature).

3.2. Alignment

For the purposes of training the speech transformer and creating perceptual test stimuli (see Section 5), it is necessary to time-align orthographically identical sentences of the dysarthric (source) and the normal (target) speech at the frame level. Let $S_{i=1\ldots N}$ and $T_{j=1\ldots M}$ represent the phoneme sequence of the source and the target speakers' rendition of the same text, respectively. We let S be the *template phoneme sequence*, prescribing the "correct" sequence of phonemes.

As expected with dysarthric speech, T often has insertions, deletions, or substitutions when compared to S. In order to best match phonemes S_i with T_j, we perform a dynamic time warping (DTW) algorithm on *phoneme feature vectors* associated with the phoneme classes. We use four one-dimensional, continuous variables as features: Voicing, Manner, Place, and Height. A multidimensional scaling has shown that these features cluster similar phonemes closely.

The DTW path's starting and ending points are constrained to be at the lower left and upper right corners of the DTW matrix, coinciding with leading and trailing pauses, that is $S_1 = S_N = T_1 = T_M = /.pau/$. The local constraints are set to allow repetition and skipping of up to 2 target phonemes at a time. Once the optimal alignment A is obtained, we additionally calculate the final distance between aligned phonemes S_i and $T_{A(i)}$ using the associated phoneme feature vectors.

Next, we take a closer look at the results and modify T according to the following algorithm:

1. If there is a single target phoneme associated with a single source phoneme, then it is either the same phoneme or a substitution, depending whether their phoneme feature distance is zero or greater than zero. No modification is necessary in this case.
2. If more than one source phoneme is associated with a single target phoneme, we designate the source phoneme with the lowest distortion to the target as a match, and the others as deletions. For deletions, we retrieve corresponding target phonemes from anywhere within the database as follows: First, we look for the phoneme in its original left and right context in any target sentence in the database. If none is found, we then look for left context only, right context only, and finally, as a last resort, we retrieve the phoneme in any context.
3. If there is a target phoneme that is not associated with any source phonemes, it is treated as an insertion and skipped.

3.3. Training and Transformation

To assemble training and evaluation sets, we collect all acoustic features of aligned phonemes, linearly deleting and repeating target features as necessary to match the number of source features

per phoneme. Due to the particular structure of the speech corpus, we use sentences 1–37 for training and sentences 38–74 for evaluation (see Section 2).

We use a Gaussian Mixture Model (GMM) to model the probability distribution of the source spectral features x as the sum of Q normal distributions with mean vector μ, diagonal covariance matrix Σ, and prior probability α

$$p(x) = \sum_{q=1}^{Q} \alpha_q \, \mathrm{N}(x; \mu_q, \Sigma_q)$$

We have trained several GMMs, some in supervised mode, and others in unsupervised mode. In supervised mode, we specify 17 different ways of partitioning the set of all phonemes into Q different classes, ensuring the uniqueness of each class and the completeness of the partitioning as a whole. For each class, we estimate μ_q and Σ_q of one mixture component by fitting the data to a normal distribution and letting α_q equal the frequency of occurrence. Then we perform a least-squares linear regression between x and the corresponding, aligned target spectral features y for each class and store the result in W_q and b_q.

In the unsupervised case, we estimate the GMM parameters with 1, 2, 4, 8, and 16 mixture components via the EM algorithm and perform a probabilistic least-squares regression to map x to y [7].

During transformation, the predicted spectral features $\hat{y}$ are calculated via a piece-wise linear probabilistic function

$$\hat{y} = \sum_{q=1}^{Q} h_q(x) \, (W_q(x) + b_q)$$

where h represents the posterior probabilities. Using a simple spectral distortion measure between y and $\hat{y}$ as an evaluation criterion, a four-component GMM trained in unsupervised mode performed best.

4. SPEECH STIMULI FOR EVALUATING INTELLIGIBILITY

4.1. Stimulus Types

Nine types of speech are created for evaluating various factors that may affect intelligibility (the term "normal" refers to non-dysarthric speech):

A. Original normal waveform,
B. LPC-synthesized representation of the normal waveform,
C. Original dysarthric waveform,
D. LPC-synthesized representation of the dysarthric waveform,
E. "Hybrid" waveform containing the spectral-envelope characteristics of the normal waveform and the prosodic (F_0, energy, and timing) characteristics of the dysarthric waveform,
F. "Hybrid" waveform containing the spectral-envelope characteristics of the dysarthric waveform and the prosodic characteristics of the normal waveform,
G. "Hybrid" waveform containing vowels, liquids, and glides (VLG) from the normal waveform and other non-approximant consonants (non-VLG) from the dysarthric waveform,
H. "Hybrid" waveform containing VLG from the dysarthric waveform and non-VLG from the normal waveform, and
I. Transformed waveform created from the baseline speech transformation system.

LPC-synthesized representations of the original waveforms are created for comparison with the hybrid and speech transformation waveforms, in order to be able to better separate effects caused by the underlying signal processing representation from effects caused by intended spectral or prosodic modifications that rely on this signal processing. The LPC order of 24 has been chosen as a compromise between having fewer features (and thus a more compact representation) and better-quality synthesis (from more detailed spectral modeling).

The hybrid waveforms containing spectral-envelope characteristics from one speaker and prosodic characteristics from another speaker are created to evaluate the effect of spectral and prosodic aspects of speech on word intelligibility, and to simulate the best possible short-term spectral speech transformation system.

The hybrid waveforms containing VLG phonemes from one speaker and non-VLG phonemes from the other speaker are created in order to (1) assess the intelligibility of a hypothetical speech transformation algorithm that modifies only formants and leaves non-formant regions of the speech signal untouched, and (2) assess the impact on intelligibility of spectral discontinuities at the boundaries between VLG and non-VLG phonemes.

The waveforms containing the baseline speech transformation system are created to evaluate not only how well standard technology performs on such a task, but also to see how closely the results from this technology match the case of the spectral-envelope characteristics of the normal speaker and prosodic characteristics of the dysarthric speaker. Large differences in these results might point to a weakness in the speech transformation algorithm or its implementation, the feature representation, or the quality or quantity of training data.

5. PERCEPTUAL LISTENING TEST

A perceptual test was employed to evaluate the word-level intelligibility of the original, LPC-synthesized, hybrid, and converted speech. This test has a structure similar to that proposed by Kent *et al.* [5] and used by Menéndez-Pidal *et al.* [4] for measuring the intelligibility of speech.

Subjects took this test using a graphical user interface that presented the stimuli and word choices, and then recorded the responses. The sentence structure was displayed on the screen, with a list of four word options each at the location of the first noun, the verb, and the second noun. The word options were not displayed until after the stimulus had been heard. A subject then selected the three words that were heard from the list of options by clicking on one word in each list. There were two stimuli at the beginning of the test that served to familiarize the subject with the procedure; responses from this familiarization stage were not recorded.

This test was taken by 18 individuals (10 males and 8 females), none of whom had known hearing problems. The stimuli consisted of the first 36 sentences in the evaluation set (sentences 38 through 73) spoken by speakers LL and/or JP from the Nemours database, each modified in the nine ways specified in Section 4. These 324 sentence-level stimuli were presented twice in total, resulting in 36 stimuli presented to each subject. Each subject listened to the same (random) sentence ordering but a different modification of each sentence, so that sentence position had no effect on the relative intelligibility results of the modifications. The stimuli were played to each subject over loudspeakers in a quiet room.

Type of Speech	Percent Corr.	Number Corr.	Total Words
Normal Speech Waveform	99	203	206
Normal Speech, LPC Synthesized	93	192	206
Dysarthric Speech Waveform	68	141	206
Dysarthric Speech, LPC Synthesized	68	141	206
Hybrid, Normal Spectrum and Dysarthric Prosody	87	179	206
Hybrid, Dysarthric Spectrum and Normal Prosody	75	155	206
Hybrid, Normal VLG and Dysarthric Non-VLG	75	155	206
(VLG word options)	72	78	108
(non-VLG word options)	79	77	98
Hybrid, Dysarthric VLG and Normal Non-VLG	73	150	206
(VLG word options)	69	74	108
(non-VLG word options)	78	76	98
Baseline Speech Transformation System	67	139	206

Table 1. Results of perceptual experiment, showing the type of stimulus, the percent correct (percent intelligible), number of correctly identified words, and total number of words for that stimulus.

The words presented as the closed-form alternatives for each word choice were derived from the original word alternatives in the Nemours database. Modifications included (a) limiting the number of choices to four for each word and (b) creating an equal number of cases in which vowels, liquids, and glides were varied in the word alternatives (VLG forms) and in which non-approximant consonants (non-VLG forms) were varied in the word alternatives. For example, a VLG form contains the words {bin, Ben, bean, ban}, and a non-VLG form contains the words {bin, tin, din, pin}.

6. RESULTS

Because of the difficulty of separating voicing (a phonemic, rather than prosodic, characteristic) from F_0 (a prosodic characteristic), five cases were removed evaluation. These were the cases in which a subject's response could be influenced by only a change in voicing and in which there was a difference in voicing production between the source and target speaker for the same word.

Results of the experiment are shown in Table 1. It can be seen that LPC re-synthesis degrades intelligibility for speaker JP from 99% to 93%, but that such resynthesis has no effect on the intelligibility of speaker LL (68%). Normal spectral characteristics with dysarthric prosody result in intelligibility of 87%, while dysarthric spectral characteristics with normal prosody result in intelligibility of 75%. Modifying only VLG regions of speech does not yield intelligibility comparable to modifying all spectral components. The baseline speech-transformation system has intelligibility of 67%, demonstrating that much more can be done for improving the short-term spectral speech transformation techniques toward a goal of 87% intelligibility.

Results of a planned-comparison one-tailed *t*-test [8] are shown in Table 2. Both prosody and spectral effects are significant at the 5% level, but only spectral effects are significant at the 1% level.

Comparison	Diff. (%)	t-test	p value
LPC effect: A+C *vs.* B+D	2.5	1.232	0.117
Spectral effect: E *vs.* D	17.7	5.952	0.000
Prosodic effect: F *vs.* D	6.6	1.824	0.043
Disruption effect: (G+H)/2 *vs.* E	12.3	3.771	0.001
Baseline transformation: I *vs.* D	-0.8	-0.200	0.422

Table 2. Results of planned-comparison one-tailed significance testing, with stimulus types indicated by the corresponding letters from the list in Section 4.

7. CONCLUSION

In conclusion, we have found that it is theoretically possible to improve the word-level intelligibility of a dysarthric speaker from 67% to 87% by modifying the short-term spectral content on a frame-by-frame basis. However, there is a large discrepancy between the intelligibility of the baseline speech transformation system and what should should be possible by modification of only spectral characteristics. Finally, it is important to fully address both VLG and non-VLG classes of speech simultaneously; thus, a speech transformation system should not address only those parts of speech containing formants.

These results offer no generality to other speakers or conditions, and in fact we expect these results to be highly speaker-dependent. However, the *methods* used here should be applicable to any dysarthric speaker, and further experiments with other speakers and transformation systems are planned.

8. REFERENCES

[1] Fried-Oken, M. and Bersani, H. A. Jr., *Speaking Up and Spelling It Out*, Paul H. Brookes Publishing Co., 2000.

[2] Electronic Speech Enhancement, Inc., "The Speech Enhancer," http://www.speechenhancer.com/.

[3] Laures, J. S. and Weismer, G., "The effects of a flattened fundamental frequency on intelligibility at the sentence level," *Journal of Speech, Language, and Hearing Research*, vol. 42, pp. 1148–1156, Oct. 1999.

[4] Menéndez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., and Bunnell, H. T., "The nemours database of dysarthric speech," *Proc. of ICSLP 1996*, pp. 1962–1965, Oct. 1996.

[5] Kent, R.D., Weismer, G. Kent, J.F., and Rosenbek, J.C., "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, 1989.

[6] Sjölander, K. and Beskow, J., "Wavesurfer – an open source speech tool," *Proc. of ICSLP 2000*, vol. IV, pp. 464–467, Oct. 2000, Beijing.

[7] Stylianou, Y., Cappé, O., and Moulines, E., "Continous probabilistic transform for voice conversion," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[8] Winer, B.J. and Brown, D. R and Michels, K.M., *Statistical Principles in Experimental Design*, McGraw-Hill, Inc., New York NY, 1991.