INTERSPEECH 2009 BRIGHTON

Classifying clear and conversational speech based on acoustic features

Akiko Amano-Kusumoto, John-Paul Hosom, Izhak Shafran

Center for Spoken Language Understanding
Department of Science & Engineering, Oregon Health & Sciences University
20000 NW Walker Road, Beaverton, OR 97006, USA
{akusumoto, hosom, zak}@cslu.ogi.edu

Abstract

This paper reports an investigation of features relevant for classifying two speaking styles, namely, conversational speaking style and clear (e.g. hyper-articulated) speaking style. Spectral and prosodic features were automatically extracted from speech and classified using decision tree classifiers and multilayer perceptrons to achieve accuracies of about 71% and 77% respectively. More interestingly, we found that out of the 56 features only about 9 features are needed to capture the most predictive power. While perceptual studies have shown that spectral cues are more useful than prosodic features for intelligibility [1], here we find prosodic features are more important for classification.

Index Terms: binary classification, acoustic features, decision tree classifier, multilayer perceptron,

1. Introduction

A speech recognition system is sensitive to changes in speaking style. As the users changes their speaking style to hyperarticulated speech, the recognition error may increase [2]. This unusual speaking style, collectively referred to as Clear speech (or CLR) here, can also occur in a variety of situations such as when speakers are talking to a hard-of-hearing listeners, or communicating in a noisy environment [3, 4]. The intelligibility of CLR is known to be higher than that of conversational (or CNV) speech, spoken as in daily communication with a colleague [3, 5]. The acoustic characteristics that are known to be significantly different between CLR and CNV speech include longer phoneme duration (or slowed speaking rates), longer and more frequent pauses, larger vowel spaces for the lax vowels, greater F0 fluctuations, often released stop bursts, and increased spectral energy in the 1000–3000 Hz range [6, 5].

If a classification stage prior to a speech recognition system can determine which speaking style a specific user is talking in, it is possible to switch to a different speech recognition system that is adapted to that speaking style. Similar to the neutral/Lombard effect classification for the two-stage recognition system [7], we propose to build the speech-style classifier using a subset of acoustic features from CLR and CNV speech in this study. The accurate classification of speaking style can not only help improve speech recognition but also help tailor strategies in a spoken dialog system. As a second goal, we will determine which features are relevant to classify the two speaking styles, CLR and CNV speech. Whether the speaking style is classified as CLR or CNV speech has, until now, been based on perception, examining if the average intelligibility of one type of speaking style is higher than the other [1]. The proposed classification algorithm may be useful to filter out non-significant CLR speech (which does not represent the characteristics of CLR speech) from a study, independent of human perception.

Previously, researchers investigated the relationship between acoustic features and speech intelligibility [8], speaker variability in different speaking styles [9], and the benefit of CLR speech for different populations [10]. Researchers examined the primary factor in determining the improved intelligibility of CLR speech by finding the correlation between acoustic-phonetic characteristics and speech intelligibility [8]. From a study by Hazan and Markham [8], a measure of long-term average spectrum was obtained over the following frequency band regions: 500–3000 Hz, 1000–2000 Hz, 500–2000 Hz, and 1000–3000 Hz. The only significant correlation with word intelligibility was found for the total energy in the 1000–3000 Hz region (male speakers $N = 15$, adult listeners $N = 15$, $r = 0.803$, $p < 0.001$). The word duration was also found to be significantly correlated with word intelligibility (male speakers $N = 15$, $r = 0.672$, $p = 0.006$). Even though these results tell us which features are expected to be relevant to a CLR/CNV classification, the correlation method examined a single feature instead of a combination of acoustic features.

Kain *et al.* [1] investigated which features contribute to the improved intelligibility of CLR speech by modifying an aspect of CNV speech to adopt CLR speech features. The results showed that the spectrum and duration information contributed to the improved intelligibility of CLR speech, while F0 and long-term energy fluctuation did not. The acoustic features that contribute to the improved intelligibility of CLR speech may, however, be different from acoustic features that are relevant to classify the two speaking styles.

In this study, we examine features that are shown to be different between CLR and CNV speech in terms of both prosodic and spectral features [6, 5]. Features are selected based on the information gain and pruned decision trees. We describe the data structure and data analysis method in Section 2, experiment using the decision tree classifier in Section 3, and experiment using the multilayer perceptron in Section 4.

2. Data Description

2.1. Data Structure

We used the OGI CLR-speech corpus [1] from three speakers (1 male and 2 females), which consists of 70 syntactically and semantically correct sentences [11]. They are phonetically balanced sentences, containing 7–10 words per utterance. Acoustic features in 210 utterances (70 utterances $\times$ 3 speakers) in each speaking style were analyzed, resulting in 420 feature vectors

This work was supported in part by NSF grant 0826654.

in total.

The input ($\mathbf{x_1}$, ... $\mathbf{x_n}$) is an N-dimensional acoustic feature vector ($N = 56$). The output is the class $C = 1$ for speaker's intention of producing CLR speech, and $C = -1$ for the speaker's intention of producing CNV speech. In the following section, the method to extract acoustic features and the summary of 56 attributes are described.

2.2. Data Analysis

Duration: All sentences were annotated and segmented automatically using forced alignment [12]. Total vowel durations, durations of the last vowel in the sentence, total consonant durations, the longest vowel duration, the longest consonant duration, mean stop burst durations (/b, d, g, p, t, k/), and total pause durations were measured per sentence and divided by the sentence duration. Burst count and pause count are the number of occurrences of bursts and pauses per sentence. Consonant-vowel duration (CVD) ratios were computed by dividing the duration of the consonants /b, d, g, p, t, k, f, v, s, z, m, n/ by the following vowel duration and averaging over the sentence [13].

The ten features we selected are: (1) vowel duration, (2) final vowel duration, (3) consonant duration, (4) maximum vowel duration, (5) maximum consonant duration, (6) CVD ratio, (7) stop burst duration, (8) stop burst count, (9) pause duration, and (10) pause count.

Fundamental frequency (F0): The fundamental frequency (F0) was extracted by taking the inverse of the distance between two consecutive glottal-closure instances (GCI) using the software Praat [14]. The F0 values (in Bark) were averaged over all vowel regions. The range was obtained by taking the difference between maximum and minimum F0 values over the entire sentence.

The two features we selected are: (11) vowel F0 mean, and (12) vowel F0 range.

Formant frequency: First and second formant trajectories and formant bandwidths were extracted using the Snack Sound Toolkit (`http://www.speech.kth.se/snack`) in the vowel regions. The formant values were taken from the middle of the vowel. Formant information was measured for the following 37 features: the mean F1 and mean F2 values of seven vowels /iː/, /ɪ/, /u/, /ɛ/, /æ/, /ʌ/, and /ɑ/ and corresponding bandwidths (BW), and the mean distance between F1 and F2 frequencies of these seven vowels. The mean distance between F1 of /iː/ and F1 of /æ/ (for the F1 range), and the mean distance between F2 of /u/ and F1 of /iː/ (for the F2 range) were included in order to estimate the vowel space. Formant frequencies were converted to the Bark scale.

The thirty seven features we selected are: (13) F1 range, (14) F2 range, (15) F1–F2 distance of vowel /iː/, (16) F1–F2 distance of vowel /ɪ/, (17) F1–F2 distance of vowel /u/, (18) F1–F2 distance of vowel /ɛ/, (19) F1–F2 distance of vowel /æ/, (20) F1–F2 distance of vowel /ʌ/, (21) F1–F2 distance of vowel /ɑ/, (22) mean F1 of /iː/, (23) mean F2 of /iː/, (24) mean F1 of /ɪ/, (25) mean F2 of /ɪ/, (26) mean F1 of /u/, (27) mean F2 of /u/, (28) mean F1 of /ɛ/, (29) mean F2 of /ɛ/, (30) mean F1 of /æ/, (31) mean F2 of /æ/, (32) mean F1 of /ʌ/, (33) mean F2 of /ʌ/, (34) mean F1 of /ɑ/, (35) mean F2 of /ɑ/, (36) mean F1 BW of /iː/, (37) mean F2 BW of /iː/, (38) mean F1 BW of /ɪ/, (39) mean F2 BW of /ɪ/, (40) mean F1 BW of /u/, (41) mean F2 BW of /u/, (42) mean F1 BW of /ɛ/, (43) mean F2 BW of /ɛ/, (44) mean F1 BW of /æ/, (45) mean F2 BW of /æ/. (46) mean F1 BW of /ʌ/, (47) mean F2 BW of /ʌ/, (48) mean F1 BW of /ɑ/, and (49) mean F2 BW of /ɑ/.

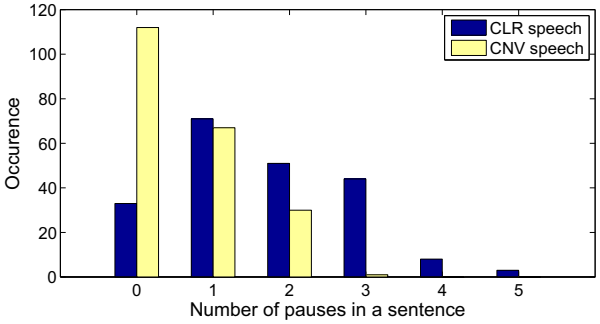

Figure 1: Histogram of the pause count

Energy: We examined the root-mean square (RMS) energy for vowels and consonants separately on energy-normalized sentences. The RMS energy range was computed by taking the difference between maximum and minimum RMS energy values. The (CV) energy ratios were calculated by dividing the RMS energy of the consonant /b, d, g, p, t, k, f, v, s, z, m, n/ by the RMS energy of the following vowel and converted to the dB scale [13].

The three features we selected are: (50) vowel RMS energy range, (51) consonant RMS energy range, and (52) CV energy ratio.

Spectrum: The long-term average spectrum (LTAS) of an energy-normalized sentence was calculated in four frequency bands (500–3000 Hz; 1000–2000 Hz; 500–2000 Hz; 1000–3000 Hz), measured in dB [8].

The four features we selected are: (53) mean spectrum between 500 and 3000 Hz, (54) mean spectrum between 1000 and 2000 Hz, (55) mean spectrum between 500 and 2000 Hz, (56) mean spectrum between 1000 and 3000 Hz.

2.3. Results of the Data Analysis

We performed a two-tailed, paired t-test ($p < 0.05$) for each feature to determine a statistical difference between CLR and CNV speech for that feature. The p values and significance of selected features are shown in Table 1. Eighteen attributes out of 56 (3 prosodic features: F0 range, vowel RMS energy, CV ratio, and 15 spectral features) did not have significant differences in their means between CLR and CNV speech. In particular, the LTAS in all four bands (53)–(56) was not shown to be significantly different, unlike a previous study [5]. The reason might be because of the speakers' characteristics; it was shown in [9] that different speakers employ different strategies to produce CLR speech.

As an example of acoustic features, Figure 1 shows the histogram of the pause count (attribute 10) in CLR and CNV speech, which has the highest information gain (0.2046) among the 56 features. The total number of pauses was increased in CLR speech from 130 to 352 instances. The number of stop burst consonants (attribute 8) showed an increase in CLR speech from 957 to 983 instances. As shown in [6], the stop consonants at the final word position are often released in CLR speech.

The information gain (IG) was calculated in each attribute by

$$\mathbf{I}(\mathbf{A}, \mathbf{Y}) = \mathbf{H}(\mathbf{A}) - \mathbf{H}(\mathbf{A}|\mathbf{Y}) \quad (1)$$

where $\mathbf{H}(\mathbf{A})$ is the entropy of attribute $\mathbf{A}$, and $\mathbf{H}(\mathbf{A}|\mathbf{Y})$ is the conditional entropy of attribute $\mathbf{A}$ given $\mathbf{Y}$. The information

Table 1: 10 best attributes in each prosodic and spectral feature group. The total of 20 attributes are shown with the mean values of CLR and CNV speech, p-values (degree of freedom in parentheses) from a two-tailed, paired t-test. Information gain (IG) and its rank are indicated as well as the rank in each prosodic and spectral feature group.

Num.	Attribute	CNV	CLR	p value (df)	IG	Rank	Feat. group w/ rank
10	Pause count	0.6190	1.6762	8.1e-028 (209) *	0.2046	1	Pros.1
9	Pause duration	0.0144	0.0459	1.4e-024 (209) *	0.1833	2	Pros.2
11	Vowel F0 mean	1.7768	1.7515	4.7e-003 (209) *	0.0874	3	Pros.3
32	Vowel F1 mean /ʌ/	4.5103	4.7738	2.6e-007 (191) *	0.0703	4	Spec.1
21	F1-F2 distance /ɑ/	3.7442	3.0258	3.3e-019 (122) *	0.0613	5	Spec.2
2	Final vowel duration	0.0717	0.0590	5.0e-028 (209) *	0.0585	6	Pros.4
5	Max consonant duration	0.0774	0.0694	3.2e-009 (209) *	0.0570	7	Pros.5
12	Vowel F0 range	1.3608	1.4360	0.1232 (209)	0.0489	8	Pros.6
46	Vowel F1 BW of /ʌ/	109.1470	97.8330	6.3e-003 (191) *	0.0390	9	Spec.3
30	Vowel F1 mean of /æ/	5.9077	6.8737	7.8e-014 (104) *	0.0341	10	Spec.4
19	F1–F2 distance /æ/	6.2097	5.4758	1.5e-009 (104) *	0.0311	11	Spec.5
20	F1–F2 distance /ʌ/	6.7361	6.3660	2.3e-009 (191) *	0.0301	12	Spec.6
52	CV energy ratio	-11.8714	-11.7694	0.6495 (209)	0.0273	13	Pros.7
6	CVD ratio	0.7999	0.7208	2.0e-002 (209) *	0.0224	14	Pros.8
51	Consonant RMS energy range	0.1144	0.1083	2.2e-005 (209) *	0.0216	15	Pros.9
50	Vowel RMS energy range	0.0913	0.0916	0.8590 (209)	0.0207	16	Pros.10
24	Vowel F1 mean of /ɪ/	4.4167	4.5888	0.0033 (131) *	0.0198	17	Spec.7
23	Vowel F2 mean of /iː/	13.9635	14.4636	7.4e-023 (101) *	0.0181	18	Spec.8
25	Vowel F2 mean of /ɪ/	12.3511	12.7778	3.4e-007 (131) *	0.0178	19	Spec.9
49	Vowel F2 BW of /ɑ/	188.4554	162.2370	1.0e-003 (122) *	0.0176	20	Spec.10

gain of the 10 best attributes in each prosodic and spectral group, for a total of 20 attributes, are listed in Table 1.

3. Experiment 1: Decision Tree Classification

The machine learning algorithm was implemented using the software Weka (ver. 3.6.0) [15]. The first experiment for feature selection was carried out using the decision tree algorithm with pruning. The training and testing was performed with 10-fold cross-validation, and was repeated 10 times with different partitions, resulting in 100 tests in total. In the training set, one fold was used for pruning, and the rests were for growing the tree. The threshold for the confidence factor was set to 5% for the tree pruning. The confidence factor indicates the percentage of values in the training set classified correctly by that path of the tree [16]. The minimum number of instances per leaf was set to 2. Five datasets, (1) all 56 attributes and (2) 9 attributes based on the pruned tree, (3) 10 best attributes based on IG, (4) 10 best attributes based on IG in the prosodic feature group, (5) 10 best attributes based on IG in the spectral feature group were examined.

The accuracy on each dataset from the decision tree algorithm are summarized in Table 2. The accuracy did not change much with and without pruning, but the size of the tree was reduced. The results from the pruned tree ($cf_{th} = 5\%$) showed that nine attributes,

(10) Pause count, (9) Pause duration,
(11) Vowel F0 mean, (32) Mean F1 of /ʌ/,
(21) F1–F2 distance of /ɑ/, (2) Final vowel duration,
(5) Max consonant duration, (12) Vowel F0 range,
(46) F1 BW of /ʌ/,

were relevant for classification with 74.17% accuracy, which was the best case using the decision tree algorithm. While 6 of the relevant features were prosodic features, 3 spectral features were included. A previous study [1] showed prosodic features, including F0, energy, and pausing, had little contribution to the improved intelligibility of CLR speech. It was unexpected that the features needed to classify speaking style are different from the features that are important for speech intelligibility.

4. Experiment 2: Multilayer Perceptron

The second experiment was carried out to determine whether a multilayer perceptron (MLP) algorithm would work better than the decision tree algorithm. The MLP is a feedforward artificial neural network that maps N-dimensional input data to a set of output classes. Each unit in each layer is represented as a perceptron,

$$y_i = f(\sum_{j=1}^{m} w_{ij} x_j + b_0) \quad (2)$$

where b_0 is the bias term. Learning weights w_{ij} are determined through backpropagation during training. The weight update equation is

$$w_{ij}(t+1) = w_{ij}(t) + \eta \Delta_j(t) y_i(t) + \alpha[w_{ij}(t) - w_{ij}(t+1)]. \quad (3)$$

In this experiment, a sigmoidal function was used for the activation function. The learning rate η and momentum α were 0.2 and 0.2, which were determined from preliminary experiments. The training and testing was performed with 10-fold cross-validation, and 10% of training set was held out for a cross-validation set. The training was terminated if the validation set error became worse 20 times in a row or maximum training time was reached (500 epochs). The entire process was repeated 10 times with different partitions, resulting in 100 tests in total. The number of units in the single hidden layer used in the experiment is listed in Table 2

The results from the MLP are summarized in Table 2 for each dataset. The best accuracy of 78.74% was obtained using

Table 2: Classification accuracies in percent (standard deviation in parentheses) of the decision tree algorithm with and without pruning and the multilayer perceptron algorithm.

Dataset	Decision Tree without pruning (%)	Decision Tree (%) ($cf_{th} = 5\%$)	MLP (%)	Num. hidden units
All 56 attributes	70.83 (6.89)	71.19 (6.53)	77.17 (5.87)	278
9 best attributes	**74.24 (7.26)**	**74.17 (6.85)**	76.14 (6.64)	44
10 best attributes	73.31 (7.38)	74.02 (7.04)	**78.74 (6.24)**	49
10 best in prosodic group	72.52 (6.43)	72.38 (7.04)	72.90 (6.93)	49
10 best in spectrum group	65.57 (6.80)	65.10 (6.80)	66.79 (6.33)	51

the 10 best attributes determined from the IG. MLP classifiers consistently outperformed decision tree classifiers for all sets of attributes and the improvement was found to be statistical significant (a two-tailed, paired t-test, $p < 0.05$) when all attributes were used.

In the decision tree algorithm, the classification accuracy did not improve with more than the 9 best attributes. The results from the MLP also showed the accuracy with the 10 best attributes (78.74%) was better than the one with all 56 attributes (77.17%). The difference between the prosodic and spectral feature group was about 7.28% for the decision tree, and 6.11% for the MLP, which indicates the importance of prosodic features. Twenty-six out of 41 spectral features (63.4%) had a significant mean difference between CLR and CNV speech, as opposed to 80.0% of prosodic features. The change in spectral features between the two speaking styles might not be as dramatic as in prosodic features, even though the spectral features are more important for intelligibility [1]. The smaller set of identified prosodic features could also be incorporated into parsimonious models to improve speech recognition [17].

In this study, the acoustic features may include incorrect phoneme duration, F0 values, and formant values, caused by automatic feature estimation. Even with these errors, the results in this work showed 78.74% classification accuracy. With all automatic feature extraction, the classification system could be implemented in real time.

5. Conclusions

In conclusion, we were able to obtain 78.74% classification accuracy using a subset of acoustic features from CLR and CNV speech. The number of features required for the classification was as low as 9 features with 76.14% accuracy. While prosodic features (i. e. F0, energy contour, and pause) were not important for the improved intelligibility of CLR speech [1], they (F0 mean and F0 range) were shown to be relevant for the classification task. The results indicate that the features that are important for speaking style classification and for intelligibility may be different.

Differences in speaking styles are not necessarily only in acoustics, but may also exist in language complexity. Combining higher-level language information with acoustic information may lead to better classification accuracy.

6. References

[1] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom, "Hybridizaing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," *Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2308–2319, 2008.

[2] E. Shriberg, E. Wade, and P. Price, "Human-machine problem solving using spoken language systems (sls): Factors affecting performance and user satisfaction," in *Proceedings of the DARPA speech and natural language workshop*, 1992, pp. 49–54.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech." *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, 1985.

[4] S. Oviatt, G.-A. Levow, M. MacEachern, and K. Kuhn, "Modeling hyperarticulate speech during human-computer error resolution," in *Proceedings of ICSLP*, vol. 2, Philadelphia, PA, 1996, pp. 428–2238.

[5] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *Journal of the Acoustical Society of America*, vol. 15, no. 1, pp. 362–378, 2004.

[6] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 29, pp. 434–446, 1986.

[7] H. Boril, P. Fousek, and H. Hoge, "Two-stage system for robust neutral/lombard speech recognition," in *Proceedings of Interspeech*, 2007, pp. 1074–1077.

[8] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *Journal of the American Academy of Audiology*, vol. 116, no. 5, pp. 3108–3118, 2004.

[9] S. H. Ferguson, "Talker differences in clear and conversatoinal speech: Vowel intelligibility for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2365–2373, 2004.

[10] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 259–271, 2002.

[11] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silberger, G. E. Urbanek, and M. Weinstock, "IEEE Recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.

[12] J. P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, pp. 352–368, 2009.

[13] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Communication*, vol. 24, pp. 211–226, 1998.

[14] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (Ver. 4.3.14)," 2005, http://www.praat.org.

[15] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2005.

[16] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[17] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, Invited Paper*, 2003, pp. 147–154.