Improved Classification of Speaking Styles for Mental Health Monitoring using Phoneme Dynamics

Keng-hao Chang[1], *Howard Lei*[2], *John Canny*[1]

[1]Computer Science Division, University of California, Berkeley, CA, USA
[2]International Computer Science Institute, Berkeley, CA, USA
{kenghao@cs, hlei@icsi, jfc@cs}.berkeley.edu

Abstract

This paper investigates the usefulness of segmental phoneme-dynamics for classification of speaking styles. We modeled transition details based on the phoneme sequences emitted by a speech recognizer, using data obtained from a recording of 39 depressed patients with 7 different speaking styles - normal, pressured, slurred, stuttered, flat, slow and fast speech. We designed and compared two set of phoneme models: a language model treating each phoneme as a word unit (one for each style) and a context-dependent phoneme duration model based on Gaussians for each speaking style considered. The experiments showed that language modeling at the phoneme level performed better than the duration model. We also found that better performance can be obtained by user normalization. To see the complementary effect of the phoneme-based models, the classifiers were combined at a decision level with a Hidden Markov Model (HMM) classifier built from spectral features. The improvement was 5.7% absolute (10.4% relative), reaching 60.3% accuracy in 7-class and 71.0% in 4-class classification.
Index Terms: speaking styles, language model, phoneme, duration

1. Introduction

This paper investigates automated classification techniques to recognize different human speaking styles for detection of mental illness. Previous work has shown that there are remarkably strong cues to mental illness in short samples of the voice. Pitch variation, pause, and information of glottal cycles are discriminative features in differentiating subjects with and without mental illness (90% accuracy with approximated 5-minute samples of speech) [1, 2]. These cues are evident in severe forms of illness, but it would be most valuable to make earlier diagnoses from a richer feature set. Furthermore, there is a semantic gap between the low-level voice features identified by speech processing technology and the diagnostic cues detected by trained practitioners. That is, practitioners develop skills and expertise in recognizing abnormal speaking styles listed in the mental status exam (Table 1) [3]. We believe that by mimicking how practitioners (or by capturing the knowledge and skills of practitioners) recognize abnormal speaking styles, it is possible to build a speech diagnostic system to detect early symptoms of mental illness [4].

We can apply methods used in emotion recognition to classify different speaking styles [5, 6]. For example, suprasegmental features such as prosody at the utterance level describe the overall textural characteristics presented by each speaking style [5]. Nonetheless, these suprasegmental features may fall short in capturing local regularities of phonology in each speaking style. For example, *pressured* speech, often characterized as "fast, virtually nonstop, seemingly driven, and hard to interrupt" [3, 4], expresses rapid phoneme transitions. On the other hand, *slurred* speech presents slower phoneme transitions due to poor pronunciation, which causes certain phonemes to be recognized less frequently than in normal speech.

Table 1: Speech Descriptors in Metal Status Exam

Category	Patterns
Rate of speech	slow, rapid
Flow of speech	hesitant, long pauses, stuttering
Intensity of speech	loud, soft
Clarity	clear, slurred
Liveliness	pressured, monotonous, explosive
Quantity	verbose, scant

The goal of this study is to explicitly model the phoneme transitions at a local, segmental level for categorizing speaking styles. We implemented the classification of 7 different speaking styles - normal, pressured, slurred, stuttered, flat, slow, and fast speech - in the framework of language modeling and duration modeling at the phoneme level. We selected and labeled the 7 styles because they were the most prominent styles utilized by practitioners in our previous exploration study [4].

We brought in techniques used in language identification, especially those exploiting the phonology difference between languages [7, 8]. Indeed, we can treat speaking styles as different "languages". Unlike the case of real world languages differing in the basic phoneme sets, different speaking styles share the same set of phonemes, a property that causes difficulty for classification. However, phoneme frequencies may differ between speaking styles. Zissman [8] has shown that language modeling at the phoneme level is effective in identifying the differences.

Language modeling represents the probability of "transition" from a phoneme to the other, but it does not necessarily capture how a phoneme "stays" as the same phoneme. Therefore, we model the durations of phonemes to depict the aspect. Duration modeling is simply the diagonals of transition matrices in Hidden Markov Models (HMMs) obtained from acoustic modeling, for which we are interested in their alternations in speaking styles. Orthogonally, we may map language modeling to the off-diagonal values. Incidentally, the durations of phonemes have also been found to be useful in the identification of slow speech [9] - a manifestation of depression.

We used a CMU Sphinx-3 speech recognizer trained with ICSI meeting corpus [10] to output phoneme and its duration sequences. Then, we built a n-gram language model on

the phoneme sequences of each speaking style and used a maximum-likelihood classifier rule, which predicts the speaking style of an unknown utterance based on the language model likelihood. For phoneme durations, we built a Gaussian model for each phoneme and its context (left and right phonemes). Following the same fashion as language models, we calculated the likelihood of an utterance by multiplying the likelihood of each phoneme duration appearing in the sequence given its context.

In addition, we investigated a proper way of user normalization. Since our study covers 39 subjects with diverse speaking styles, taking into account individual differences (a human practitioner must also do this) is important to accurately recognize individual speaking styles. For language modeling, we normalized at the model level - a general cross-user model was adapted to a user-specific model with back-off interpolation. This method resembles the technique used for automatic speech recognition (ASR) in a specific domain. To improve search accuracy, ASR language models that are larger and more general are often adapted to smaller, domain-specific models [11]. For duration modeling, we normalized at the feature level - we normalized the duration of a phoneme by subtracting the duration average of a user. This method is similar to the one used for spectral features, i.e. cepstral mean normalization [12]. For comparison, we also designed a baseline classifier with spectral features, using Mel-frequency cepstral coefficients (MFCCs) and Hidden Markov Models (HMMs), an approach that has provided some promising results in speaking style recognition [13], speaker identification [14], and emotion recognition [15].

The paper is presented as follows. Section 2 describes the speech database we used. Section 3 explains the procedure for training and testing language models, duration models, and HMMs with MFCCs. Section 4 describes the experimental results and section 5 concludes the paper.

2. Speech Database

We recorded one-on-one therapy sessions spanning 10 months, in which 39 depressed patients and their health practitioners consented to participate in the recordings. The recording took place in doctors' offices, which were quiet and have microphone arrays (Acoustic Magic Voice Tracker at the sampling rate of 44 kHz) placed 2 feet away and on the side of the patients and doctors. This setup maintained recording quality while minimizing interference to the sensitive nature of therapies. Because abnormal speaking styles appeared occasionally in the natural conversation, we segmented the recordings and labeled the utterances that match the 7 speaking styles of interest - normal, pressured, slurred, stuttering, flat, slow, and fast speech. We consulted the practitioners so that the labeling conform to the perception of the practitioners [4], where we adapted a master/apprentice model [16] to transfer the domain knowledge. Altogether, we compiled 1297 utterances that last a total of 3.9 hours, with utterance averages of 10.7 seconds. We obtained 66.9 minutes of normal speech, 38.1 minutes of pressured speech, 16.1 minutes of slurred speech, 34.6 minutes of stuttering speech, 49.00 minutes of flat speech, 9.9 minutes of slow speech, and 16.7 minute of fast speech. The data imbalance is due to the fact that some speaking styles happen more frequently than others. Among the 7 speaking styles, some styles share similar characteristics. For example, pressured and fast speech both contain rapidness of speech, but pressured speech also includes more animated texture (e.g. higher intensity). Both flat and slow speech presents slowness and paused speech, but flat speech has an additional low animation feature (e.g. low vocal energy and pitch variation). Each utterance is associated with only one class label, and in our experiments, we downsampled the speech data to 16 kHz.

3. Language and Duration Modeling of Phoneme Transitions

We made use of a CMU Sphinx-3 speech recognizer trained with the ICSI meeting corpus [10] to output phoneme and duration sequences. Using MFCCs with delta and acceleration coefficients, the speech recognizer generates phoneme sequences with 4000 tied senones, each equipped with 32 Gaussian mixtures.

First, we built a general n-gram language model P_s for speaking style s using the phoneme sequences labeled as speaking style s. With the 7 models, the language model classifies an unknown utterance with phone sequence W as speaking style $\hat{s}$ based on the speaking style with the highest likelihood, i.e.

$$\hat{s} = arg\max_s \mathcal{L}_s(W) \tag{1}$$

$$\mathcal{L}_s(W) = \sum_t \log P_s(w_t|h_t) \tag{2}$$

where w_{t-1} and w_t are consecutive phonemes observed in the phone sequence W, with history $h_t = w_{t-1}, \cdots, w_{t-n+1}$. We used the CMU statistical language model toolkit [17] for experimentation, and we applied Witten Bell discounting technique [18] for unseen events. For user normalization, we first built a user-specific n-gram model $P'_{s,u}$ using a separate held-out dataset, which includes phoneme sequences of user u labeled as style s. Because the model is small, it may fail to capture the general speaking-style trends. Therefore, we back-off interpolated the model $P'_{s,u}$ with the larger, general model P_s (3). Finally, the intermediate model $\tilde{P}_{s,u}$ was linearly interpolated with the style-specific model P_s, generating the final version of the user-adapted model $P_{s,u}$ (4). Using the same held-out data that we used to train $P'_{s,u}$, we calculated the optimal interpolation weights via the Expectation-Maximization algorithm [19].

$$\tilde{P}_{s,u}(W_u) = \begin{cases} P'_{s,u}(w_t|h_t) & if P'_{s,u}(w_t|h_t) \geq T \\ \lambda P_s(w_t|h_t) & otherwise \end{cases} \tag{3}$$

$$P_{s,u}(W_u) = \alpha \tilde{P}_{s,u}(W_u) + \beta P_s(W_u) \tag{4}$$

where T is an empirical threshold, the back-off coefficient λ was calculated so that $\tilde{P}_{s,u}(w_t|h_t)$ sums to one, and W_u is an utterance of user u.

For modeling phoneme durations, we built Gaussian models Q_s for each speaking style s. In particular, we constructed a Gaussian for each phoneme w, using the observed durations of the phoneme w and its context (left and right phonemes) that appear consecutively in utterances labeled as style s. Following the same fashion of n-gram modeling, we calculated the likelihood of an utterance by multiplying the likelihood of each phoneme duration appearing in the sequence $D = \{d_t\}$ given its context, i.e.

$$\mathcal{M}_s(D, W) = \sum_t \log Q_s(d_t|w_t, h_t, h'_t) \tag{5}$$

$$\hat{s} = arg\max_s \mathcal{M}_s(W) \tag{6}$$

where $h_t = w_{t-1}, \cdots, w_{t-n+1}$ and $h'_t = w_{t+1}, \cdots, w_{t+n-1}$. For user normalization, we adapted the duration feature before the training and testing was performed, where the new duration

Table 2: Confusion matrix of 2-gram language modeling at phoneme level with user normalization. Rows represent the test utterances, and columns represent the predicted speaking styles. *Norm* stands for normal speech, *Pres* stands for pressured speech, *Slur* stands for slurred speech, and *Stut* stands for stuttering speech. For easier interpretation, the confusion matrix is normalized in each row (i.e. a column does not sum to one). We also list the F_1 harmonic measure between precision and recall for each class in the bottom row. The weighted average accuracy was 53.2% (0.73).

	Norm	Pres	Slur	Stut	Flat	Slow	Fast
Norm	0.65	0.11	0.02	0.05	0.11	0.02	0.04
Pres	0.18	0.56	0.07	0.03	0.02	0.00	0.15
Slur	0.03	0.04	0.69	0.05	0.08	0.04	0.07
Stut	0.19	0.11	0.10	0.43	0.08	0.03	0.06
Flat	0.17	0.07	0.12	0.04	0.50	0.04	0.06
Slow	0.11	0.26	0.06	0.07	0.21	0.28	0.01
Fast	0.33	0.21	0.05	0.04	0.02	0.00	0.35
F_1	0.61	0.56	0.49	0.52	0.55	0.27	0.35
ROC	0.71	0.73	0.76	0.73	0.73	0.68	0.75

d'_t was computed by subtracting the averaged phoneme duration of user s across all phonemes and styles, i.e.

$$d'_t = d_t - \frac{\sum_{W \in u,t'} d_{t'}}{N} \quad (7)$$

4. Experimental Results

4.1. Language Modeling Classifier of Phoneme Sequence

For evaluation, we performed a 10-fold leave-subject-out cross-validation. That is, we trained models on the utterances of 35 subjects, tested on the utterances of the remaining 4 subjects, and iterated. Because we are interested in building a model that can be generalized to new users, the same user does not appear in both the training and test utterances. In addition, this user-based break-up will better reflect the effect of user normalization. Moreover, we did not want to create a "back-door" to recognizing speaking styles if the distribution of speaking styles is different between speakers. The language model could actually classify speakers who have different mixtures of speaking styles. We computed weighted average accuracy for speaking styles $\{s\}$ to compare the performance of classifiers, i.e.

$$AvgAccuracy = \frac{\sum_s Accuracy_s * TotalDuration_s}{\sum_s TotalDuration_s} \quad (8)$$

For language models, we first tested the general cross-user models P_s. The weighted average accuracy was 27.7% with 2-gram models. Then, we performed user normalization. We did back-off interpolation to create user-adapted models $P_{s,u}$ using 1/5 of utterances of test user u as a held-out set, and tested the model with the remaining 4/5 of utterances of the user. The threshold T for back-off was set to 0.1 empirically. The accuracy significantly improved to 53.2% and the details are given in Table 2. On average, the size of the phoneme sequences for building each general model P_s was 130019 phonemes, whereas the user-adapted models $P_{s,u}$ were based on a held-out set of 2592 phonemes.

Table 2 shows that the system did not recognize slow and fast speech as well as the others ($F_1 < 40\%$) . This may be partially due to the fact of imbalanced data, where each of these styles occupies less than 10% of the entire dataset. In addition, because some speaking styles share similar characteristics, these speaking styles are likely to be miss-classified. For example, some fast speech was classified as pressured speech, slow speech was confused as flat speech, etc. There is also a tendency that speaking styles with varied speech rate (e.g. fast and pressured speech) were confused with normal speech. This brings us to a hypothesis that language modeling does not model speaking styles with varied speech rate very well. Lastly, one thing surprised us was that slow speech was confused with pressured speech. Again, this may due to the imbalance of data.

Table 3: F_1 measures of 2-gram and 1-gram duration modeling with user normalization. Weighted average accuracies were 27.5% and 28.9% (0.58 and 0.69)

2-gram	Norm	Pres	Slur	Stut	Flat	Slow	Fast
F_1	0.39	0.25	0.17	0.12	0.29	0.04	0.16
ROC	0.53	0.56	0.65	0.51	0.65	0.57	0.58
1-gram	Norm	Pres	Slur	Stut	Flat	Slow	Fast
F_1	0.30	0.48	0.11	0.12	0.28	0.22	0.25
ROC	0.55	0.85	0.63	0.52	0.76	0.78	0.74

4.2. Gaussian Classifier of Phoneme Duration Sequence

To evaluate duration models, we also applied a 10-fold leave-subject-out cross validation. Table 3 lists the F_1 measures of a 2-gram duration modeling with user normalization. The weighted average accuracy was 27.5%, where user normalization only gave about 4% improvement. We have also experimented with division-based normalization, instead of the subtraction-based normalization in equation (7), but normalization by subtraction achieved better improvement. Note that the normalization applied here is a feature-based normalization and the system did not use any held-out data, which is likely to be a major factor for poor improvement.

The result of 1-gram duration modeling is also given in Table 3, in which the F_1 measures for speech rate-varied classes were better than those in the 2-gram duration model. For example, F_1 of slow speech increased from 4% to 22%, fast speech increased from 16% to 25%, and pressured speech went up from 25% to 48%. Because a 1-gram model does not consider phoneme context (left or right phonemes), the predicted likelihood is purely based on the duration of phonemes, which should favor classes with varied speech rate. In addition, since 1-gram duration models are inclined to classify based on speech rate, normal speech has worse accuracy because other classes such as slurred or stuttered speech have the same speech rate.

4.3. Hidden Markov Model Classifier with Spectral Features

For comparison, we designed a classifier based on spectral features, using HMMs and 39-dimension MFCCs, which include delta and acceleration coefficients. We made use of the Hidden Markov Toolkit (HTK) [20] for training and testing. We trained a general HMM for each speaking style, and a maximum likelihood-based classification decision was made. For user normalization, we applied maximum likelihood linear regression (MLLR) method [20] to adapt the mean values of each Gaussian, using the 1/5 held-out utterances of a test user. A 1-state, 1 diagonal Gaussian mixture HMM achieves 54.5% weighted average accuracy (Table 4). The MLLR-based normalization was significant, providing a 30% boast in accuracy.

Table 4: F_1 measures of 1-state HMM classifier with MFCCs and user normalization. Weighted average accuracy was 54.5%. (0.77)

	Norm	Pres	Slur	Stut	Flat	Slow	Fast
F_1	0.66	0.66	0.40	0.45	0.52	0.34	0.44
ROC	0.76	0.67	0.83	0.62	0.82	0.87	0.83

Table 5: Classification accuracy for different classifiers

Classification Method	Avg Accuracy
2-gram language model	53.2%
HMM with MFCCs (spectral)	54.5%
Combination of the 2 classifiers for 7 classes	61.7%
Combination of the 2 classifiers for 4 classes	72.1%

The result shows that, although the HMM classifier using MFCCs (54.5%) performs slightly better than the language model at the phoneme level (53.2%), it is an encouraging result that the local phonological dynamics provide useful information for classifying speaking styles. We also combined the classifiers at the decision level, and the result is given in Table 5. The phonology information improved the weighted average accuracy to 60.3% (a 5.7% absolute and 10.4% relative improvement over the HMM classifier). Finally, we tested the classifier on fewer classes. We did this by ignoring the classes which might suffered from data imbalance, and performed a 4-class classification between normal, pressured, stuttered and flat speech. The accuracy reached 71.0%.

5. Conclusions

We investigated speaking style recognition using language models and Gaussian duration models with phoneme sequences. The results showed that local regularities of phonology provide useful information to improve speaking style recognition. Combined with a HMM classifier using short-term spectral features, phoneme transition information was able to achieve the recognition accuracy of 60.3% for 7-class and 71.0 % for 4-class classification. Nonetheless, the combined decision showed only modest improvement, which may be due to the fact that these two classifiers were based on similar features, although they are at different levels. The spectral features describe the sound textures that people perceive. Assisted by HMMs to model the change of the features, to some extent the HMM classifier is able to depict the phonology transition. On the other hand, phoneme-based language modeling abstracts the change of spectral features by looking at the high-level phoneme sequences, which were emitted using acoustic models. Nonetheless, we argue that language modeling better mimics people's higher level perception (i.e. whether certain phoneme is pronounced correctly in slurred speech).

These lead to topics that we should further explore. First, we could combine some orthogonal prosodic features such as pitch and glottal timings with the the current classifiers in order to improve the performance [5]. Second, we will leverage the existing features and classifiers to build a mental health monitor, a model that predicts the health level given the human voice. During the collection of the speech database, we also collected the health ratings of patients from the doctors, so we can perform a correlation or classification analysis according to the ratings. In particular, we are interested in putting the speaking styles as a latent variable for predicting mental health levels.

Finally, we plan to relax the requirement that we use a held-out dataset for user normalization. If we consider applying this model to real world users, it would be somehow cumbersome to ask an expert to label some utterances of each new user. A possible way to do this is to generate some rough estimate of speaking styles, use it as an held-out dataset, and re-estimate until the result converges.

6. References

[1] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," in *IEMBS*, 2004.

[2] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *Am J Psychiatry*, vol. 154, pp. 4–17, 1997.

[3] P. T. Trzepacz and R. W. Baker, *The Psychiatric Mental Status Examination*. Oxford University Press, 1993.

[4] K. Chang, M. K. Chan, and J. Canny, "Analyzethis: Unobtrusive mental health monitoring by voice," in *CHI Extended abstracts*, 2011.

[5] M. Eskenazi, "Trends in speaking styles research," in *Eurospeech*, 1993.

[6] P. Juslin and K. Scherer, *Vocal expression of affect*. Oxford University Press, 2005, ch. 3, pp. 65–135.

[7] V. W. Zue, T. J. Hazen, and T. J. Hazen, "Automatic language identification using a segment-based approach," in *Eurospeech*, 1993, pp. 1303–1306.

[8] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[9] F. Martinez, D. Tapias, J. Alvarez, and P. Leon, "Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition," in *Eurospeech*, 1997.

[10] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at icsi," in *ICASSP*, 2003.

[11] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communications*, vol. 42, pp. 93–108, 2003.

[12] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effects of the communication channel in auditory-like analysis of speech," in *Eurospeech*, 1991.

[13] K. Ravikumar, R.Rajagopal, and H.C.Nagaraj, "An approach for objective assessment of stuttered speech using MFCC features," *ICGST International Journal on Digital Signal Processing, DSP*, vol. 9, pp. 19–24, 2009.

[14] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *ICASSP*, 2008.

[15] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *ICSLP*, 2004.

[16] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, 1997, ch. 3. Principles of Contextual Inquiry, pp. 41–66.

[17] R. Rosenfield and P. Clarkson, "Statistical language modeling using the cmucambridge toolkit," in *Eurospeech, 5th European Conference on Speech Communication and Technology*, 1997.

[18] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, 1991.

[19] W. Xu and A. Rudnicky, "Language modeling for dialogue system," in *ICSLP*, 2000.

[20] S. Young, "The htk hidden markov model toolkit: Design and philosophy," University of Cambridge: Department of Engineering, Cambridge, UK, Tech. Rep., 1994.