Automatic Estimation of Perceived Sincerity From Spoken Language

Brandon M. Booth, Rahul Gupta, Pavlos Papadopoulos,
Ruchir Travadi, Shrikanth S. Narayanan

Signal Analysis & Interpretation Laboratory (SAIL)
Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering
University of Southern California, Los Angeles, CA, USA

http://sail.usc.edu

Abstract

Sincerity is important in everyday human communication and perception of genuineness can greatly affect emotions and outcomes in social interactions. In this paper, submitted for the INTERSPEECH 2016 Sincerity Challenge, we examine a corpus of six different types of apologetic utterances from a variety of English speakers articulated in different prosodic styles, and we rate the sincerity of each remark. Since the utterances and semantic meaning in the examined database are controlled, we focus on tone of voice by exploring a plethora of acoustic and paralinguistic features not present in the baseline model and how well they contribute to human assessment of sincerity. We show that these additional features improve the performance using the baseline model, and furthermore that conditioning learning models on the prosody of utterances boosts the prediction accuracy. Our best system outperforms the challenge baseline and in principle can generalize well to other corpora.
Index Terms: Behavioral Signal Processing (BSP), computational paralinguistics, speech assessment, sincerity, challenge

1. Introduction

Speech production is a complex process whose subtleties have profound impacts on our ability to communicate effectively. Changes in acoustic production, gestures, posture, and facial expressions all affect others' perceptions of our intent, mood, and mental state. Small variations in these expressive modalities can have acute implications in many common and high-stakes social situations such as relationship formation, business deals, and political negotiations [1].

This year's INTERSPEECH Sincerity Challenge focuses on assessing perceived sincerity exclusively from acoustic modalities. This emphasis is important for understanding and predicting outcome from communication in expressivity-constrained settings such as telephony. Several studies have examined sincerity in spoken dialog and there is overwhelming consensus that prosodic cues play a fundamental role in sincerity perception not only as rated by humans [2][3][4][5], but also as interpreted at a neural level [6]. Prosody serves many functions, helping to differentiate questions from statements and convey attitude and affect, but these conveyances are more difficult to differentiate from sarcasm and irony reliably in the absence of context [5]. This challenge limits the potential impact of this confusion by focusing on utterances pertaining to apologies and by fixing the word choice. It further provides both monotonic and non-monotonic utterances for many of the speakers in the corpus allowing for further exploration of the relative effects of prosodic variance on sincerity.

Speaking rate and the prominence provided by pauses between phonemes are also important cues for assessing genuineness [5]. For example, Anolli et al. show that among other paralinguistic differences, speech rate is faster when expressing irony (one form of insincerity) [7]. This challenge also provides fast and slow variants of utterances to enable exploration of the role of this paralinguistic feature in sincerity assessment relative to other acoustic dynamics.

Although some researchers have studied sincerity indirectly through their efforts in irony recognition, very little work seeks to understand it directly and how it is differentiated from other types of dishonesty [8]. In this paper, we aim to address two research questions:

1. Which acoustic and paralinguistic features are helpful in assessing sincerity?
2. Given a classification of prosodic style and no additional context, can the prediction accuracy of genuineness be improved?

We propose a set of features supplemental to the challenge's baseline ComParE features that improves performance when using the baseline model and also propose a mixture-of-experts framework that exploits latent human perception biases of sincerity based on prosodic style for further estimation gains.

2. Corpus and Baseline System

The corpus and learning problem are explained in detail in the challenge paper [8], but we provide a brief summary here. The dataset consists of audio files (655 training, 256 test) from 32 different speakers each containing a single utterance from a set of six prespecified apologies. Ten of these speakers are withheld as test data and the task is to predict sincerity scores on a normalized scale. For each utterance, speakers are instructed to articulate in one of four different styles: fast, slow, monotonic, or non-monotonic. The gold standard sincerity labels are averaged across a small group of annotators to mitigate annotator bias and a Spearman correlation performance metric is employed.

Table 1 shows the distribution of training data per utterance and per speech style. Each type has sufficient representation in this corpus. Table 2 shows the mean sincerity ratings per style-utterance type that have been normalized beforehand per speaker.

Style \ Utterance ID	1	2	3	4	5	6
F	23	25	20	20	21	26
S	30	29	25	30	27	32
M	31	26	28	20	21	24
L	46	29	25	36	31	30

Table 1: Distribution of training set utterances per utterance ID and per speech style across all speakers. F=fast, S=slow, M=monotonic, L=non-monotonic

Style \ Utterance ID	1	2	3	4	5	6
F	-.327	-.082	.171	.115	.142	-.115
S	-.199	.121	.210	.225	.331	.165
M	-.312	.168	.280	.125	.193	-.132
L	-.467	-.147	-.025	-.011	-.044	-.25

Table 2: Distribution of mean normalized perceived sincerity per utterance ID and per speech style across all speakers in the training set. F=fast, S=slow, M=monotonic, L=non-monotonic

The baseline algorithm uses a linear SVM regressor to estimate sincerities based on statistical functions of low-level features detailed in [9].

3. Feature Extraction

Several acoustic features have been successfully employed for humor and sarcasm detection and by proxy sincerity evaluation [3][4][5]. Based on these reports and in particular a thorough analysis of features conducive to deception detection [2], we extract the following features and describe the process in more detail below:

- Formants F1, F2, F3
- Intensity
- Arousal
- Gabor kernel convolutions of the spectrogram
- Gammatone Frequency Cepstral Coefficients (GFCC)
- i-Vectors
- Aligned Phones

The first three formants (F1, F2, F3), and intensity features are extracted using Praat [10]. We also compute arousal scores using a robust model proposed by Bone et al [11] that has been validated on multiple affective data sets. Three features (F0, intensity, HF500) are combined and normalized to produce an arousal score in the range $[-1, 1]$ via:

$$p_j = \sum_i w_i \cdot p_{i,j} \tag{1}$$

where p is the arousal, j indexes an utterance, and i indexes a feature. The per-feature arousal is given by:

$$p_{i,j} = 2F_{i,j}(f_{i,j}) - 1 \tag{2}$$

where f is a feature and F is some reference cumulative distribution function for feature i. Finally, w_i is a normalized weight scalar given by:

$$w_i = \frac{r(\boldsymbol{p}_i, \bar{\boldsymbol{p}})}{\sum_{i'} r(\boldsymbol{p}_{i'}, \bar{\boldsymbol{p}})} \tag{3}$$

Here r is the Spearman correlation function, $\boldsymbol{p}_i$ is the i-th feature's arousal score vector over all utterances per speaker, and $\bar{\boldsymbol{p}}$ is the mean of $\boldsymbol{p}_i$ over all features. In our case, the reference F function is estimated per speaker using the percentage of utterances with lower feature values, and w_i is normalized across all utterances for that speaker.

Gabor filters are physiologically inspired edge orientation detectors applied to a spectro-temporal representation of the speech signal [12]. Recent findings indicate that some neurons in the primary auditory cortex of mammals are explicitly tuned to spectro-temporal patterns [13] [14]. Gabor filters emulate these patterns to identify spectro-temporal receptive fields (STRF), which in turn act as estimates for this neurological representation of the sound stimuli. We extract Gabor based features as described in [12].

GFCCs have been successfully used in many speech-related tasks such as voice activity detection [15], automatic speech recognition [16] [17], and speaker identification [18]. GFCCs are obtained from a bank of gammatone filters, which has been proposed to model human cochlear filtering. The impulse response of a gammatone filter in the time domain is given by:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t) \tag{4}$$

where f_c is the central frequency of the filter. The constant a controls the gain, while n defines the order of the filter. Finally b is the bandwidth and increases proportional to f_c according to:

$$b = 1.019 \cdot 24.7 \cdot (4.37 \cdot f_c/1000 + 1) \tag{5}$$

GFCC features of 24 dimensions are extracted using 64 gammatone filters (of unit gain and order one), the outputs of which are temporally integrated using a Hanning window. The center frequencies of the gammatone filter bank ranges from 50 Hz to 8000 Hz and are equally distributed based on the ERB scale [19].

I-vectors are obtained using Total Variability Modeling [20], a popular framework for capturing acoustic variability by mapping variable length sequences into a fixed low-dimensional representation. The model assumes that feature vectors in an utterance are distributed according to a Gaussian Mixture Model (GMM), and that the GMM mean supervectors $\mathbf{M}_u$ for each utterance, formed by stacking all the component means in a single vector, lie along a low dimensional linear subspace $\mathbf{T}$. More specifically:

$$\mathbf{M}_u = \mathbf{M}_0 + \mathbf{T} \cdot \boldsymbol{x}_u \tag{6}$$

where $\mathbf{M}_0$ represents the global mean supervector from a Universal Background Model (UBM), and $\boldsymbol{x}_u$ is known as the *i-vector* for that utterance. Although initially proposed for speaker recognition, i-vectors have also been found to be useful in tasks involving the inference of paralinguistic information [21].

Since the utterance text is provided in the training set, we produce word level and phonetic alignments using the Kaldi ASR toolkit [22] and acoustic models based on the Wall Street Journal corpus [23]. In the testing phase, we evaluate every utterance on each of the six possible choices and use the most likely to generate word and phone alignments.

For each of these feature types, except for arousal and phone features, we compute the minimum, maximum, upper and lower quartiles, mean, median, variance, and range, and add

each as a new separate feature. From the aligned phones, we compute the mean, minimum, and maximum phone and pause durations as well as the average speech rate (phones/sec) and duration of the entire utterance. Finally, we add mean and variance statistics of $\log(F0)$ and $\Delta\log(F0)$ aggregated over utterance frames.

In total this leaves us with 3138 features supplementing the baseline set. All functionals of these features of similar types (e.g. all those computed from phones) are grouped and either included or excluded from training/testing altogether. This reduces the number of combinations necessary to explore during optimization.

4. Training and Prediction

We first evaluate our supplemental features using the baseline learning algorithm. We train a linear SVR using a leave-one-speaker-out strategy (LOSO) to preserve data independence during cross-validation. Next, we implement feature selection and redundant feature removal strategies to improve the performance of the baseline algorithm utilizing our additional features. Finally, we employ a supervised learning model to create a classifier that assigns speech style labels based on all aggregated features, and then also four regressors to output sincerity scores given a subset of the data conditioned on the speech type label. We explore a number of different feature selection and removal approaches such as Pearson/Spearman correlation, variance and mutual information thresholding, Chi-squared independence testing, and recursive feature elimination (RFE). We also test several supervised learning techniques including random forests, linear/kernelized SVM, rank SVM, k-nearest neighbors, and gradient boosting. All tests have been conducted using scikit-learn [24] and Weka [25]. In this paper we only discuss our best models and feature inclusion approaches in detail.

4.1. Measuring Performance

The goal in measuring error is to assess how well any given model can represent relevant structure in the data, so it is important that the overall performance is informative of the quality of choice of models. Rather than computing and averaging performance at every iteration during LOSO cross-validation, the baseline algorithm aggregates predictions from each unique trained model during LOSO-CV into one prediction vector for the entire training set and then reports the Spearman (rank-based) correlation. This approach requires the predictions from different models with different learned parameters to be stacked before performance is assessed making it difficult to tell whether a poor performance is due to the choice of learning algorithms (model bias) or sensitivity to splits in the data (model variance).

We hypothesize that a weighted average of the performance at each LOSO step yields a more informative measurement of the expected performance of a given model. Thus we propose the following Spearman correlation metric:

$$r = \sum_i w_i \cdot r_i \tag{7}$$

where i is the LOSO iteration index, r_i is the Spearman correlation between predicted and gold standard sincerity scores, and w_i is the proportion of CV-test data during the i-th iteration.

A proof of performance consistency using this metric during LOSO-CV is outside the scope of this paper. Several consistency proofs showing convergence in agreement between cross-validated training and testing results have been made when optimizing over the space of linear learning models [26] [27] [28]. None of these proofs address convergence when aggregating CV predictions, but those with desirable convergence properties average performance across CV iterations as we propose. Using a weighted average in our case is intuitive since the number of utterances per left-out speaker differs.

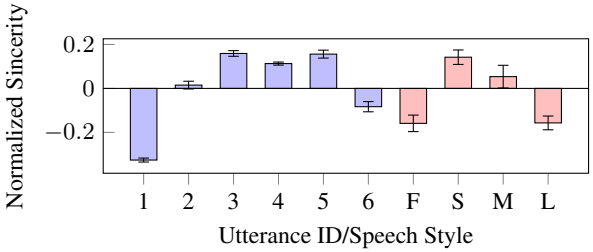

Figure 1: Average normalized sincerity rating across all subjects per utterance ID (on the left) and per speaking style (on the right) in the training set. F=fast, S=slow, M=monotonic, L=non-monotonic

4.2. Supplemental Features with the Baseline Model

In our first experiment, we reproduce the baseline algorithm and supplement the baseline feature set with our extracted features to test their advantages. Each feature is Z-normalized by subtracting the feature's mean and dividing by the standard deviation across all training samples.

4.3. Feature Selection and Redundant Feature Removal

We further refine this approach by implementing feature selection and redundant feature removal processes to reduce the dimensionality during training and eliminate noisy features. We instigate a selection policy where we include the top percentage of features for which the Pearson correlation with the target sincerity label is the highest. Among the selected features we compute pairwise Pearson correlations and, for each pair of highly correlated features, we remove the one with the lowest correlation to the target sincerity label. After these selection and removal processes we Z-normalize each feature.

Our best model using this approach selects the top 80% of features with the best Pearson correlation, and then removes the less correlated feature from each pair for which the Pearson correlation is above 0.99. A linear SVR (C=0.0001, ϵ=0.1, L2 loss) algorithm yields the highest Spearman correlation with the training set sincerity scores.

4.4. Speech Style Conditional Model

The utterance ID marginals shown in Figure 1 highlight disparities in context bias and, though this fact can be exploited for machine learning purposes, would not scale well cross corpora due to the essentially unbounded number of potential utterances. Prosodic biases are apparent in the speech style marginals which we utilize and can generalize to spoken language in any context.

We partition the data by conditioning upon the speaking style, either fast, slow, monotonic, or non-monotonic, since the corpus guarantees that every utterance fits into one of these categories. The training set provides these labels, but they are missing from the test set so we train a separate classifier to assign them. Our best speech type classifier uses a kernelized SVM

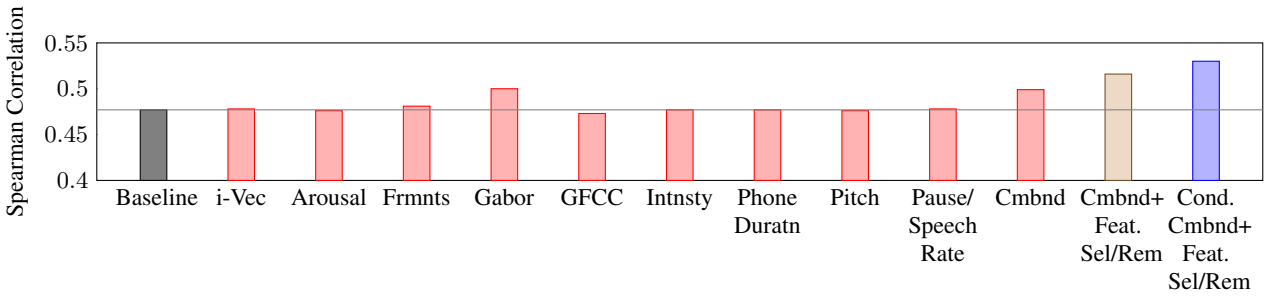

Figure 2: Spearman correlation between training set sincerity predictions and gold standard using baseline features augmented with different features. All results use the baseline learning model except the last one which uses our model conditioned on the speech style.

with our feature selection, removal, and normalization procedure.

With the output of this classifier, we partition the training data into four groups and then train four individual models (experts) to rate sincerity using linear SVR. A mixture-of-experts committee machine is used for prediction where the output from the "best" expert estimates sincerity and the best expert is selected after examining the speech style label for each incoming utterance.

Our top performing model uses a radial basis function kernel SVM (C=0.95, ϵ=0.001) to assign speech style labels and the same linear SVR settings from before for the four expert models. We arrived at these optimal C parameter values for the speech style classifier using a grid search on the interval [0.85,1.15] with 0.05 spacing and for the expert models using the interval [10^{-6},10^{-2}] with powers of ten spacing. The gamma parameter for the RBF kernel was set to $\gamma = \frac{1}{\text{\# features}}$.

5. Results and Discussion

All reported Spearman correlation results from training use our modified metric from equation (7). Figure 2 shows the Spearman correlation performance of the baseline model augmented with different sets of our supplemental features and also our best feature selection/removal model and best model conditioned on speech style.

Among all of the individual supplemental features we test, Gabor features offer the largest improvement over the baseline. Unfortunately, these features are not easily interpretable. Our best combined feature set with the baseline learning model performs comparably when trained on only baseline and Gabor features suggesting that little additional usable information is provided by all of our other supplemental features given this learning model. Therefore, Gabor features seem to be the most powerful supplemental discriminators for sincerity assessment.

Our best augmented baseline model with all combined features and dimension reduction techniques uses 8235 features which includes our Gabor, formants, i-vectors, arousal, and intensity features. Incorporating any of our other features degrades our performance when using a linear SVM. Our best model here achieves a 0.513 Spearman correlation on the training data set and a 0.614 correlation on the test set, which beats the baseline model test set correlation of 0.602. Given this performance, we consider these additional features to be important for sincerity estimation but insufficient and still far from providing a complete foundation for understanding.

Finally, our mixture of experts approach for the data conditioned on speaking style outperforms all of our other models on the training set achieving a Spearman correlation of 0.531. In theory, this generalizes very nicely to other corpora because it only requires an accurate assessment of the prosodic style of the utterance. On the test data set, however, this approach yields a 0.557 Spearman correlation performing worse than the corpus baseline model. We attribute this deficiency to our speech style classifier which achieves an accuracy of 98% on the training data using LOSO-CV, even with vigilant efforts to avoid overfitting, but which we suppose is not generalizing well. In part, this might be accredited to the categorical overlap among the corpus speech style labels thereby making differentiation between, for example, fast and monotonic speakers difficult. The statistically significant improvement this approach brings during training leaves us optimistic regarding its viability given more accurate speech style label predictions.

6. Conclusion and Future Work

In this paper we show that several paralinguistic features aid in sincerity evaluation. In particular, Gabor-filtered spectro-temporal features and careful selection and removal of redundant features contribute most significantly. We also show that learning models conditioned on the prosodic style of utterances can outperform naïve models, but require accurate style labels to achieve the performance gain.

Continued research effort on the speech style conditional model appears to be a prudent direction for further work. Next steps here entail obtaining better speech style predictions and then assessing whether the error disparity of the mixture-of-experts model is in fact due to poor style labels. One interesting approach that may improve speech style prediction is splitting these labels into mutually exclusive binary subsets (e.g. fast vs. slow, monotonic vs. non-monotonic) and training separate binary classifiers for each. Some exploration of methods for combining these results possibly including ensemble averaging or boosting might lead to accuracy improvements. Alternately, using automatic clustering with non-unique class labels may improve speech style prediction as well.

While learning models conditioned on utterances are not as generalizable and scalable, we are curious to see what performance gains could be realized using this approach. If large sincerity assessment gains can be made in this manner, then further research into the effects of interactions between prosodic and contextual cues on sincerity would be warranted.

7. Acknowledgements

We extend our sincere appreciation to the INTERSPEECH 2016 challenge organizers and also the National Science Foundation for supporting this research.

8. References

[1] G. H. Graham, J. Unruh, and P. Jennings, "The Impact of Nonverbal Communication in Organizations: A Survey of Perceptions," *Journal of Business Communication*, vol. 28, no. 1, pp. 45–62, 1991.

[2] F. Enos, "Detecting Deception in Speech," Ph.D. dissertation, Columbia University, 2009.

[3] H. S. Cheang and M. D. Pell, "Acoustic Markers of Sarcasm in Cantonese and English." *The Journal of the Acoustical Society of America*, vol. 126, pp. 1394–1405, June 2009.

[4] E. Hoicka and M. Gattis, "Acoustic Differences Between Humorous and Sincere Communicative Intentions," *British Journal of Developmental Psychology*, vol. 30, no. 4, pp. 531–549, 2012.

[5] A. Nauke and A. Braun, "The production and perception of irony in short context-free utterances," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China, 2011, pp. 1450–1453.

[6] S. Rigoulot, K. Fish, and M. D. Pell, "Neural Correlates of Inferring Speaker Sincerity from White Lies: An Event-related Potential Source Localization Study," *Brain Research*, vol. 1565, pp. 48–62, 2014.

[7] M. G. Anolli, Luigi and Ciceri, Rita and Infantino, "Irony as a game of implicitness: Acoustic profiles of ironic communication," *Journal of Psycholinguistic Research*, vol. 29, pp. 275–311, 2000.

[8] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception , Sincerity & Native Language," in *Proceedings INTERSPEECH 2016*, ISCA, San Francisco, USA, 2016.

[9] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, 2013.

[10] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," 2016. [Online]. Available: www.praat.org

[11] D. Bone, C. C. Lee, and S. Narayanan, "Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 201–213, 2014.

[12] M. R. Schaedler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal Modulation Subspace-spanning Filter Bank Features for Robust Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 131, pp. 4134–4151, 2012.

[13] A. Qiu, C. E. Schreiner, and M. A. Escabi, "Gabor Analysis of Auditory Midbrain Receptive Fields: Spectro-temporal and Binaural Composition," *Journal of Neurophysiology*, vol. 90, no. 1, pp. 456–476, 2003.

[14] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme Representation and Classification in Primary Auditory Cortex," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 899–909, 2008.

[15] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A Robust Frontend for VAD: Exploiting Contextual, Discriminative and Spectral Cues of Human Voice," in *Proceedings INTERSPEECH 2013*, ISCA, Lyon, France, 2013.

[16] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *2013 IEEE International Symposium on Circuits and Systems*, Beijing, China, 2013.

[17] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An Auditory-based Feature for Robust Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taiwan, Japan, 2009.

[18] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating Auditory Feature Uncertainties in Robust Speaker Identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, USA, 2007.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing, Fifth Edition*. Academic Press, 2003.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[21] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of Cognitive Load from Speech Using an i-Vector Framework," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, Malaysia, 2014.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, USA, 2011.

[23] J. Garofolo, D. Graff, D. Paul, and D. Pallet, "CSR-I (WSJ0) Complete LDC93S6A," DVD, Philadelphia, 1993.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, June 2009.

[26] J. Shao, "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[27] Y. Yang, "Consistency of cross validation for comparing regression procedures," *The Annals of Statistics*, vol. 35, no. 6, pp. 2450–2473, 2007.

[28] P. Zhang, "Model Selection Via Multifold Cross Validation," *The Annals of Statistics*, vol. 21, no. 257, pp. 299–313, 1993.