INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Enhancement of Automatic Oral Presentation Assessment System using Latent N-Grams Word Representation and Part-of-Speech Information

Wen-Yu Huang[1], *Shan-Wen Hsiao*[1], *Hung-Ching Sun*[1], *Ming-Chuan Hsieh*[2], *Ming-Hsueh Tsai*[2], *Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]National Academy for Educational Research, Taiwan

Abstract

The development of an automatic oral presentation assessment system is important for the educational researchers to assess and train the communication ability of school leaders. In this work, we aim at enhancing the performance of the existing pre-service school principals' presentation scoring system by including lexical information as an additional modality. We propose to use latent n-grams distributed word representations and weighted counts of part-of-speech tag to derive features from the speech transcripts in the National Academy for Educational Research (NAER) oral presentation database. We carry out two different experiments: Exp I is a binary classification task between high versus low performing speech, and Exp II is a continuous scoring on the entire dataset. In Exp I, the proposed framework achieves a competitive accuracy of 0.79, and in Exp II, by fusing this text-based system to the existing audio-video based system, we obtain a spearman correlation of 0.641 (18.05% relative improvement). The two experiments demonstrate the modeling power of our proposed framework and signify the substantial complementary information in the lexical modality while assessing the quality of an oral presentation.

Index Terms: behavioral signal processing, multimodal signal processing, educational research, natural language processing

1. Introduction

The need for developing domain-aware computational methods that can measure human behaviors quantitatively in order to perform large-scale automatic high-level subjective assessment has sparked a vast amount of ongoing interdisciplinary research effort between engineering and behavior sciences [1]. Some notable algorithmic development already exists in health-related [2, 3], art-related [4], and education-related applications [5, 6]. The goal of deriving such a computational framework is to mitigate perennial issues centered around human subjectivity, e.g., time-consuming and error-prone manual observational coding procedures, and at the same time, to provide domain experts quantitative decision-making tools that can supplement their current capability. In this work, we extend and improve upon our previous research development toward a complete automatic oral presentations assessment for pre-service school principals by incorporating lexical information in addition to the previously-proposed audio-video based system [7, 8]. This is a joint collaborative effort with researchers from NAER.

The NAER is entrusted by the Ministry of Education in Taiwan with pre-service school principals' training (certification) program. This is a yearly mandatory program that every principal candidate participates in order to be certified, and the goal is to foster well-rounded school leaders in the current climate of high demand for complex educational reforms [9, 10, 11]. As part of the program, each candidate performs a 3-minute long impromptu speech on a randomly-selected topic. Candidates are grouped into classes, and there are two *coaching* principals, i.e., senior school principals in office, served as instructors per class who are in charge of the grading and teaching. Throughout the years of implementing this program, the NAER researchers start to recognize the difficulties in sustaining an adequate pool of capable coaching principals (resulting in having the same group of instructors every year) and further realizing the subjectivity nature in the oral presentation grading process can potentially be problematic not only for this principal-ship certification but also makes it non-scalable for other important certification program, e.g., annual teacher-ship certification.

Hence, a research effort is underway to develop an automatic assessment system based on behavioral data collected during the program. The 2014 NAER pre-service school principals oral presentation database [7] include both audio-video recordings and the manual transcripts of each impromptu speech. Since the judgment of a well-perceived oral presentation depends holistically on a combination of prosodic characteristics [12, 13], non-verbal gestures [14, 15], and also the linguistic contents [16], we expect the inclusion of lexical modality can further improve the automatic assessment system. In fact, past works have demonstrated that improved recognition rate can be achieved by modeling/fusing the lexical content in tasks such as sentiment analysis [17], high-level behavior construct [18], and emotion detection [19, 20] due to the fact of multimodal nature of human's expressive behaviors.

Our core lexical feature extraction framework utilized in this work, i.e., latent n-grams distributed word representation served as the document vector (a vector representing an entire oral presentation), is inspired by the recent success of utilizing distributed word representations for text information retrieval [21], latent semantic analysis [22, 23], and sentiment detection [24]. We construct this document vector by concatenating low-dimensional representations of different *n*-grams of distributed word vectors in order to capture varying-length contextual information embedded within such word-vectors. Aside from the word-based information, we additionally include n-grams information on part-of-speech tags and further encode it at the document-level using term-frequency inverse-word frequency. In this work, we carry out two different experiments: Exp I) binary classification between the extreme sets of data, and Exp II) complete continuous scoring for the entire database. In Exp I, we achieve an unweighted average recall of 0.79, i.e., the average of 0.74 and 0.84 for the original and rank-normalized version of the final score respectively. In Exp II, by fusing the lexical modality to the audio-video based continuous scoring system, the improved system obtains an average Spearman

http://dx.doi.org/10.21437/Interspeech.2016-400

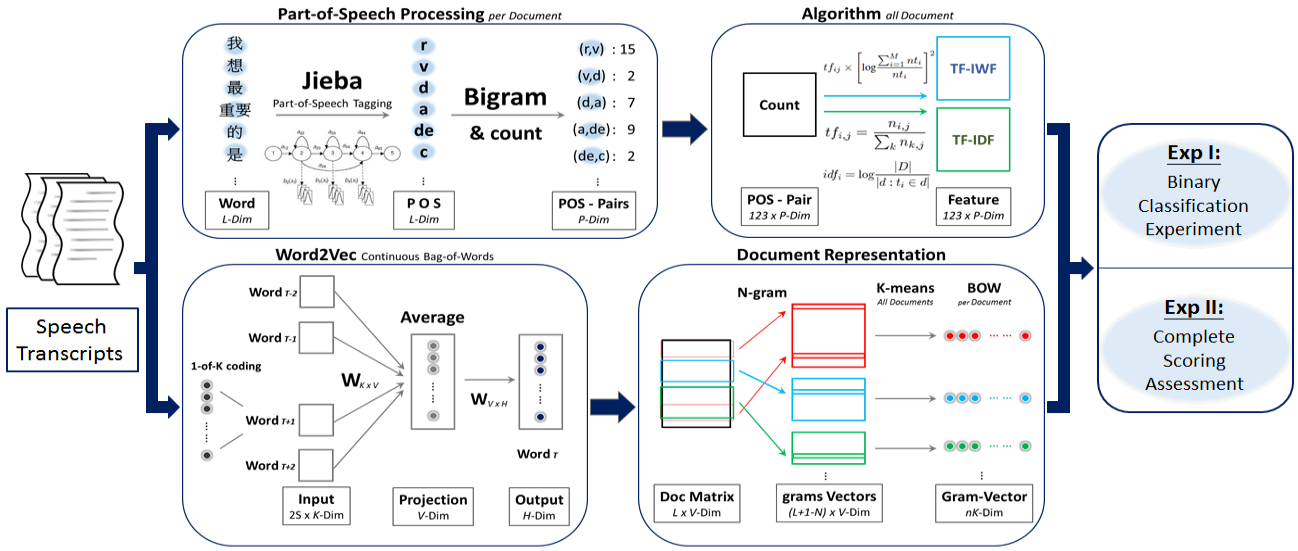

Figure 1: *It depicts the complete workflow of our lexical feature extractions: the top portion shows the part-of-speech based feature extraction, and the bottom portion shows the latent n-grams distributed word representations. We further present their accuracies in the binary classification experiment and the continuous scoring assessment in this work.*

correlation of 0.642 (0.663 and 0.621 for original and rank-normalized score respectively), which is 0.098 improvement absolute (18.05% relative) over the audio-video based framework [8]. Our experiment demonstrates both the promising discriminative power of our proposed framework and the lexical modality's complementary information benefiting a more holistic and robust assessment of an individual oral presentation.

The rest of the paper is organized as the following: section 2 describes the research methodology including database and lexical feature extraction, section 3 details the two experiments and results, and finally section 4 concludes with future work.

2. Research Methodology

2.1. The NAER Oral Presentation Database

The NAER oral presentation database consists of the 200 principal condidates' audio-video data recorded at the premise of NAER during the 2014 certification program; the manual transcriptions of these speeches are also available as part of the database. Since there are only 128 oral presentations that have scores labeled by both of their corresponding coaching principals, and 5 of the presentations are spoken in local dialects (prohibiting the transcription to be carried out in Mandarin Chinese), the final dataset utilized in this work is the 123 documents of oral presentation transcripts, where each of them is associated with two coaching principals' scoring sheets.

Each speech is graded on seven dimensions of attributes [7]: content, structure, word, etiquette, prosody, and timing. Each principal candidate's impromptu speech presentation score, which counts 5% toward the final grade that each candidate receives at the end of the program, is the summation of the average score given by the two instructors in this corresponding class on these seven dimensions. In this work, we focus on deriving automatic scoring of this final score by incorporating lexical information. Lastly, since each coaching principal can have a different dynamic range of scoring, we further employ rank label normalization [25] to mitigate this issue. Eventually, our dataset-of-interest in this work consists of 123 documents of speech transcripts each with two labels (*original* and *rank-normalized* total score). The range of the original total scores is a real value from 0 to 100, and 0 to 1 for rank-normalized scores. The Cronbach's alpha for the original and rank-normalized total score computed for the entire 123 documents are 0.44 and 0.55, respectively; the moderate inter-evaluator agreement further reinforces the need for developing such an objective assessment system.

2.2. Lexical Feature Extraction

We construct document vector at the level of an oral presentation to quantify the textual information within each transcript such that it can be utilized in the development of this system. There are two types of document-level features used in this work (Figure 1): latent n-grams word representation (*nGrams-Vec*) and weighted n-grams counts of part-of-speech tags (*pos*); each of which will be described in section 2.2.1 and 2.2.2 respectively. Lastly, since there is no space between the written characters in Chinese language, we first utilize Jieba toolbox[1] to perform automatic *word* segmentation (a *word* may consists of multiple Chinese characters) and further remove punctuations before carrying out lexical feature extraction.

2.2.1. Latent N-Grams Word Representation

Figure 1 bottom portion depicts the extraction procedure of our latent n-grams word-based features. The procedure essentially involves training a distributed word representation model and using k-means based bag-of-word encoding on n-grams average of distributed word-vectors to obtain a final document vector. In the following section, we will briefly describe each component.

The main idea behind using distributed word presentation is to use vector representations on *words* through the construction of probabilistic neural network language model [26]. The word2vec model introduced by Mikolov *et al.* is an effective and efficient methodology to achieve such a representation [27, 28]; in fact, word2vec model belongs to the class of hierarchical probabilistic neural language model [29]. The framework aims at modeling the implied relationship between a word and its surrounding words in a given corpus. There are two different methods to construct word2vec models [30], i.e., continuous bag-of-word (CBOW) model and skip-gram model. We use CBOW model due to its better performance in this work com-

[1] https://github.com/fxsjy/jieba

Table 1: *Exp I classification accuracies measured by unweighted average recall (UAR): D denotes TF-IDF, W means TF-IWF, and GV is the nGram-Vec. The subscript indicates the number 'N' in n-gram. The 's' means the feature derived from concatenating different n-grams together along with train-set univariate feature selection. The fusion framework is a logistic regression based decision-level fusion between part-of-speech feature Ws and nGrams-Vec feature GVs.*

	Word (w)				Part-of-Speech (pos)					*nGrams-Vec*					Optimized
Feature	D_{1w}	D_{2w}	W_{1w}	W_{2w}	D_{1pos}	D_{2pos}	W_{1pos}	W_{2pos}	Ws_{pos}	GV_2	GV_3	GV_4	GV_5	GV_s	Fusion
Original	0.57	0.63	0.66	0.63	0.53	0.66	0.57	0.76	0.74	0.67	0.58	0.59	0.46	0.76	0.74
Rank-Norm.	0.68	0.57	0.61	0.57	0.57	0.69	0.74	0.66	0.71	0.57	0.65	0.63	0.58	0.81	0.84
AVG.	0.62	0.60	**0.63**	0.60	0.55	0.68	0.65	0.71	**0.72**	0.62	0.62	0.61	0.52	**0.78**	**0.79**

pared to the skip-gram model. As demonstrated in Figure 1, the architecture of CBOW consists of input, projection, and output layers. Given a total number of K unique words in a corpus, with each word encoded by 1-of-k coding, the matrix $W_{K\times V}$ is the word-vector matrix (K words $\times V$ dimensions); V denotes the number of hidden semantic dimensions. When training the word-vector for a target word in a sentence, CBOW looks for $2S$ surrounding word context in order to compute the probability of the appearance of that particular target word. The output layer, instead of 1-of-k coding, consists of Huffman coding (i.e., hierarchical softmax) in order to reduce the computational cost. The projection layer matrices, $W_{K\times V}$ and $W_{V\times H}$, are updated recursively using backward error propagation. Effectively, the CBOW is a neural network that embeds a target word's semantic information onto a V-dimensional space through the process of predicting the current word using its surrounding words. The word2vec model is trained on a large background corpus to obtain robust estimation of these distributed word-vectors.

With the distributed word representation, we can then transform each speech transcript in the NAER database into a $L \times V$ document matrix, where the L is the total number of the words arranged in sequence for the document. Then, in order to further characterize the semantics of phrases (or simply a sequence of words), we compute n-gram average of the V-dimension word-vector. Lastly, we use k-means clustering approach to characterize the latent semantics of these n-grams *phrases* and perform bag-of-word encoding to obtain a document vector of size K. However, since the latent information can vary with different n, we perform this k-means bag-of-word encoding on different n, i.e., ranging from bi-gram to penta-gram, and finally we concatenate each output to form the final latent n-grams word representation document vector, termed *nGram-Vec* in this work.

2.2.2. Weighted N-grams Counts of Part-of-Speech

Aside from quantifying word-based information, we additionally include information about the word class, i.e., part-of-speech. It is intuitive that part-of-speech would carry information about the overall perceptual judgment of an oral presentation; for example, a proper usage of adverb or other forms of decorative can signify a better talk.

Figure 1 top portion depicts the feature extraction procedure for the weighted n-grams counts of part-of-speech. In this work, we use Jieba toolbox for automatic part-of-speech (pos) tagging. The tagging is done based on a combination of rule-based algorithm and Viterbi decoding in a hidden Markov model framework that is pre-trained on a large number of annotated corpora. Instead of the commonly-used term-frequency inverse-document frequency (TF-IDF) weighting features, we use term-frequency inverse-word frequency (TF-IWF). IWF uses inverse word frequency instead of inverse document frequency:

$$\text{TF-IWF}_{ij} = tf_{ij} \times \left[\log \frac{\sum_{k=1}^{M} nt_k}{nt_i}\right]^2 \quad (1)$$

where M is the numbers of words in the corpus, nt_i denotes the number of times a word i appears in the corpus. TF-IWF emphasizes more on the relationship between the word frequency and importance in a corpus; it has been shown to achieve better results on text categorization among small corpus [31]. In the end, we form a document vector, termed weighted n-grams counts of part-of-speech, by computing TW-IWF on both individual word-level part-of-speech tags (i.e., 52 tags) and paired-word (hence, $N = 2$ for the n-gram in this case) part-of-speech tags (i.e., 993 total).

3. Experimental Setup and Results

We setup two experiments in this paper. Experiment I involves binary classification (classifying between *high* versus *low* performing presentations) among the extreme set of data (top and bottom 20%) using our proposed lexical features. The Cronbach's alpha for this reduced set is 0.73 and 0.86 for original and rank-normalized scores respectively. The high inter-evaluator agreement among this extreme set can help eliminate concerns on whether the system is learning on the 'correct' labels; hence it provides a scheme for testing the efficacy of our technical framework. Experiment II is how the assessment would work in real life, where each of the presentation is given a score. We evaluate whether an improvement can be obtained by fusing lexical-based system with audio-video based system. The testing scheme is done via leave-one-speaker-out cross validation.

A large background corpus is required to train word2vec model described in section 2.2.1. We collect Chinese articles through web crawler from website such as Wikipedia, Yahoo News, and moedict. Our word2vec model is trained on the complete database containing 525,747,491 raw words, and the vocabulary consists of about 2 million Chinese words and 1 million non-Chinese words (hidden dimension is set at $V = 400$). K-means bag-of-word encoding is done with $k = 64$, determined empirically.

3.1. Experiment I: Binary Classification

In Exp I, we use support vector machine (linear kernel with C=1) as the classifier. The evaluation metric is unweighted average recall. Table 1 summarizes the results of the Exp I. For word-based features, we evaluate the use of various n for nGram-Vec (denoted as GV_*), TF-IDF (denoted as D_*), and TF-IWF (denoted as W_*). For part-of-speech based features, we present results on TF-IDF and TF-IWF. The numerical subscript indicates the number n used in the n-grams. The 's' means that the classifier is trained on features derived from concatenating different n-grams together along with train-set univariate feature selection. The final optimized fusion model is carried out using logistic regression on the decision scores from part-of-speech feature Ws_{pos} and nGram-Vec GV_s.

There are several points to note in the results. First of all, the best results from this lexical-only modality framework achieves an average UAR of 0.79 (0.74 and 0.84 for original

Table 2: *Exp II continuous scoring results presented in Spearman correlation:* AV-BEST *is the previous audio-video framework [8]. Subscript w and pos indicates the word-based and part-of-speech based features with their associated document vector encoding methods, the fusion technique is simple averaging.*

		Original score	Ranked score	Average
Binary$_{SVM}$	Ws_{pos} + GV_s	0.514	0.533	0.523
	AV-BEST	0.528	0.558	0.543
	AV-BEST + D_{1w}	0.455	0.457	0.456
	AV-BEST + W_{1w}	0.491	0.474	0.483
	AV-BEST + D_{2pos}	0.561	0.525	0.543
	AV-BEST + Ws_{pos}	0.558	0.549	0.553
	AV-BEST + GV_s	0.567	0.563	0.565
	AV-BEST + Ws_{pos} + GV_s	**0.663**	**0.621**	**0.641**

and rank-normalized score) in this binary classification task. The result is competitive with the best audio-video based system [8], which achieves an average UAR of 0.83 (0.85 and 0.81 for original and rank-normalized score) under identical experimental setting. Secondly, modeling of part-of-speech seems to be more effective than straightforward modeling of lexical content (see columns of *Word* and *Part-of-Speech* in Table 1). In specifics, bi-gram weighted counts of part-of-speech feature (W_{2pos}) alone achieves an accuracy of 0.71 - signifying the importance of incorporating word class information in context. In fact, by examining the feature selection output for the case of mono-gram part-of-speech, some of the important features are from the class of "d" (adverb), "ag" (use noun or adjective as an adverb to decorate a verb), "a" (adjective), and "v"(verb). Lastly, another interesting point to make is that any fixed number of grams in the computation of latent distributed word representations (nGram-Vec) alone does not work well, but the result from concatenating different numbers of grams together can improve the classification accuracies significantly. It seems to implicate that the inclusion of varying-length windows of word information is important to capture the essential characteristics for grading these presentations.

3.2. Experiment II: Continuous Scoring

In Experiment II, we focus on predicting the real-valued score for all the presentations (i.e., not limited to the 40% of the data) by fusing the lexical modality with the audio-video based framework. Hence, the continuous scoring framework is set to be identical to our previously proposed method [8] to enable straightforward fusion and fair comparison. Our audio-video based system is based on a framework, termed Binary-SVM. The Binary-SVM uses the best-distinctive subset of the NAER data to build a classifier, then for each data, z, we then compute the distance using the trained SVM as follows:

$$dist(z) = \sum_{i=1}^{n} y_i \alpha_i x_i^T z + \rho \quad (2)$$

where $y_i \in \{1, -1\}$ corresponds to the class label of each support vector, α_i is the weight parameter for each support vector x_i, and ρ is the bias term. After generating $dist$, we linearly transform it to a range of $[1, 10]$ - this normalized 'distance-to-hyperplane' is used as the assessment score for each presentation. It has been demonstrated to be superior than conventional support vector regression [8]. The fusion scheme is, hence, straightforward by summing up the scores from each modality.

Table 2 summarizes the results of the Exp II. The evaluation metric is the Spearman correlation. AV-BEST denotes the best system achieved by using audio-video information [8]. Subscript *w* and *pos* indicates the word-based and part-of-speech based features respectively each with their associated document vector encoding methods (e.g., TF-IDF and TF-IWF); GV_s is the latent n-grams distributed word representation features (section 2.2.1). The best result is obtained using the fusion of "AV-BEST + Ws_{pos} + GV_s", which achieves an average Spearman correlation 0.641 (0.664 and 0.621 for original and rank-normalized score respectively) - an absolute improvement of 0.098 (18.05% relative) over the AV-BEST. While the lexical-only features on average does not directly outperform the AV-BEST (0.523 versus 0.543), it is encouraging to see that it indeed provides substantial complementary information to the audio-video based system. Lastly, we also observe that the inclusion of both word-based and part-of-speech-based information become essential to achieve a significant boost when combing with the audio-video based system compare to using either word-based or part-of-speech based only features.

4. Conclusions

In this work, we propose to enhance the audio-video based preservice school principals' oral presentation automatic assessment system by including text information. The lexical features are derived from latent n-grams distributed word representations and TF-IWF counts of part-of-speech tag from the speech transcripts. Our experiment (section 3) demonstrates the promising modeling power of the proposed lexical features, and by fusing this extra modality, it improves the overall accuracy obtained by the audio-video based assessment system.

There are multiple future directions for this work. On the technical side, the use of distributed word representation is recently a prevalent technique for analyzing the textual content [32, 33, 34]. While the current 2014 NAER database is limited in scale, we are collecting this data yearly at the NAER and will start exploring other techniques, e.g., convolution neural network, in achieving addition improvement. Furthermore, the automatic segmentation of Chinese language into *words* and the background corpus for word2vec training both play an important role in obtaining consistent and robust results, we will further investigate techniques to improve upon these two fundamental aspects in the development of the entire system. On the side of education, with the continuing development of audio-video-lexical multimodal assessment systems, we will begin working closely with the NAER researchers to perform a larger-scale validity and usability analyses by comparing these system-based scores versus scores derived from the concurrent administrations tests for the existing training program, e.g., written test grade, essay score, which constitutes the rest of the 95% of the final grade.

5. Acknowledgments

Thanks to MOST, Taiwan (103-2218-E-007-012-MY3) and NAER, Taiwan (NAER-104-12-B-2-02-00-1-02) for funding.

6. References

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[2] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[3] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[4] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *Multimedia, IEEE Transactions on*, vol. 16, no. 6, pp. 1766–1778, 2014.

[5] J. Whitehill, M. Bartlett, and J. Movellan, "Automatic facial expression recognition for intelligent tutoring systems," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–6.

[6] J. Tepperman, S. Lee, S. Narayanan, and A. Alwan, "A generative student model for scoring word reading skills," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 348–360, 2011.

[7] S.-W. Hsiao, H.-C. Sun, M.-C. Hsieh, M.-H. Tsai, H.-C. Lin, and C.-C. Lee, "A multimodal approach for automatic assessment of school principals' oral presentation during pre-service training program," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] ——, "A computational framework toward automatic scoring of pre-service school principals oral presentation using fusion of audio-visual information," *Affective Computing, IEEE Transactions on*, under review.

[9] D. L. Keith, "Principal desirabilitiy for professional development," *Academy of Educational Leadership Journal*, vol. 15, no. 2, p. 95, 2011.

[10] T. Bush, E. Kiggundu, and P. Moorosi, "Preparing new principals in south africa: the ace: School leadership programme," *South African Journal of Education*, vol. 31, no. 1, pp. 31–43, 2011.

[11] M. Coelli and D. A. Green, "Leadership effects: School principals and student outcomes," *Economics of Education Review*, vol. 31, no. 1, pp. 92–109, 2012.

[12] O. Kang, "Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance," *Language Assessment Quarterly*, vol. 9, no. 3, pp. 249–269, 2012.

[13] L. Chen, C. W. Leong, G. Feng, and C. M. Lee, "Using multimodal cues to analyze MLA'14 oral presentation quality corpus: Presentation delivery and slides quality," in *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. ACM, 2014, pp. 45–52.

[14] M. Gentilucci and M. C. Corballis, "From manual gesture to speech: A gradual transition," *Neuroscience & Biobehavioral Reviews*, vol. 30, no. 7, pp. 949–960, 2006.

[15] D. McNeill, *How language began: Gesture and speech in human evolution*. Cambridge University Press, 2012.

[16] J. B. Hirschberg and A. Rosenberg, "Acoustic/prosodic and lexical correlates of charismatic speech," 2005.

[17] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *Multimedia, IEEE Transactions on*, vol. 10, no. 7, pp. 1342–1355, 2008.

[18] M. Black, P. G. Georgiou, A. Katsamanis, B. R. Baucom, and S. S. Narayanan, "" you made me do it": Classification of blame in married couples' interactions by fusing automatically derived speech and language information." in *INTERSPEECH*, 2011, pp. 89–92.

[19] A. Kazemzadeh, S. Lee, and S. Narayanan, "Fuzzy logic models for the meaning of emotion words," *Computational Intelligence Magazine, IEEE*, vol. 8, no. 2, pp. 34–49, 2013.

[20] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models," *Proc. CMMR*, pp. 492–507, 2012.

[21] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 101–110.

[22] J. Bjerva, J. Bos, R. van der Goot, and M. Nissim, "The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity," in *SemEval 2014: International Workshop on Semantic Evaluation*, 2014, pp. 642–646.

[23] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2177–2185.

[24] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 208–212.

[25] H. D. Kim, C. Zhai, and J. Han, "Aggregation of multiple judgments for evaluating ordered lists," in *Advances in information retrieval*. Springer, 2010, pp. 166–178.

[26] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*. Springer, 2006, pp. 137–186.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[29] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model." in *Aistats*, vol. 5. Citeseer, 2005, pp. 246–252.

[30] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[31] R. Basili, A. Moschitti, and M. Pazienza, "A text classi er based on linguistic processing," in *Proceedings of IJCAI*. Citeseer, 1999.

[32] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[33] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[34] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.