SENSEI: SPOKEN LANGUAGE ASSESSMENT FOR CALL CENTER AGENTS

Abhishek Chandel, Abhinav Parate, Maymon Madathingal, Himanshu Pant,
Nitendra Rajput, Shajith Ikbal, Om Deshmukh, Ashish Verma

IBM India Research Lab, Vasant Kunj
New Delhi-110070, India
vashish@in.ibm.com

ABSTRACT

In this paper, we present a system, called Sensei, for assessment of spoken English skills of call center agents. Sensei evaluates multiple parameters of spoken English skills, *i.e.*, articulation of sounds, correctness of lexical stress in words and spoken grammar proficiency. Sensei provides an assessment test to be taken by a call center agent (or candidate) and generates score on each of the spoken English parameters as well as a combined score. It is implemented in the form of a web application so that it can be accessed through a web browser and doesn't require any software to be installed at the client side. We describe how the individual parameters are assessed in Sensei using various speech processing techniques and the experiments conducted to evaluate these techniques. The performance is compared with assessment performed by human assessors. A correlation of 0.8 is obtained between overall score generated by Sensei and human assessors on a real life test dataset of 243 candidates which compares well with the corresponding human-to-human correlation of 0.91.

Index Terms— speech recognition, articulation, syllable stress, grammar evaluation

1. INTRODUCTION

The success of an offshore call center organization depends to a large extent upon the communication skills of its agents in the language of their customers. Hence, voice and language assessment becomes an integral part of their hiring and training processes which is presently performed by human assessors. Sensei is a comprehensive web-enabled spoken English assessment system which evaluates various parameters of spoken English, such as, pronunciation, syllable stress pattern, spoken grammar and listening comprehension. In this paper, we focus only on the first three parameters as comprehension is currently evaluated in Sensei through multiple choice questions.

A significant amount of related work has been done for Computer Assisted Language Learning (CALL) [1, 2, 3, 4, 5]. Authors in [1] have developed two language learning applications (Tball Literacy Assessment and Tactical Language Training System) for assessment of non-native speakers. The pronunciation variations were modeled in 8-dimensional articulatory feature space (jaw, lip separation, tongue tip, tongue body etc). Syllable stress was evaluated using prosodic features like duration of the syllable nucleus, fundamental frequency, energy and their slopes and ranges. Authors in [2] have developed a hand-held pronunciation evaluation device that uses the log-posterior probability of the expected phone-sequence from an Automatic Speech Recognition (ASR) to compute the goodness of pronunciation and the pitch contour to evaluate the intonation quality. Authors in [3] have proposed syllable stress evaluation techniques based on prosodic features related to fundamental frequency, duration and energy using neural networks that use contextual syllabic information and using first and second order Markov chains. Authors in [4] model pronunciation as a combination of the speaker's knowledge of the correct phonetic transcription of a written text and the speaker's ability to pronounce the phonemes of the target language correctly. Authors in [5] propose a strategy of modeling the pronunciation variations at the syllable level using different subsets of context features (like lexical stress of the syllables, their position within the word, word's identity).

Most of the systems described above focus mainly on pronunciation evaluation and a few of them focus on evaluating syllable stress. These systems often operate in a learning framework where the user is asked to pronounce a word or a phrase and a feedback is provided to the user on the spoken utterance. The user can also listen to the model pronunciation and then using the feedback he or she can improve the pronunciation. Sensei, on the other hand, is designed to comprehensively assess a candidate on spoken English skills across multiple parameters in order to determine if the person suits a given profile. It is in the form of a test which consists of various modules, *viz.*, articulation, syllable stress, grammar and comprehension, and generates a combined score as well individual scores for each of the parameters in real-time. After the test, a business decision can be made based on this score. Sensei also provides the flexibility of modifying the content for various modules, such as sentences to be read, grammatical errors to be tested and words to test syllable stress, so that it can be customized to test the spoken English skills for a specific domain or process.

Some of the other technologies that offer spoken English assessment and instruction either over the phone or over the Internet are Ordinate [6] and GlobalEnglish [7].

Rest of the paper is organized as follows. In Section 2, we briefly describe the system architecture of Sensei and its main components. Section 3 presents in detail the approaches used to evaluate articulation, syllable stress and spoken grammar parameters. Various experiments conducted to evaluate the performance of each individual module and the corresponding results are presented in Section 4. We conclude in Section 5 and acknowledge several contributors in Section 6.

2. SYSTEM ARCHITECTURE

A high level system architecture of Sensei is depicted in Fig. 1. Sensei is designed as a web application so that the assessments can be conducted remotely and hence the offshore call center organization can reach a larger talent pool with lower cost and high efficiency.

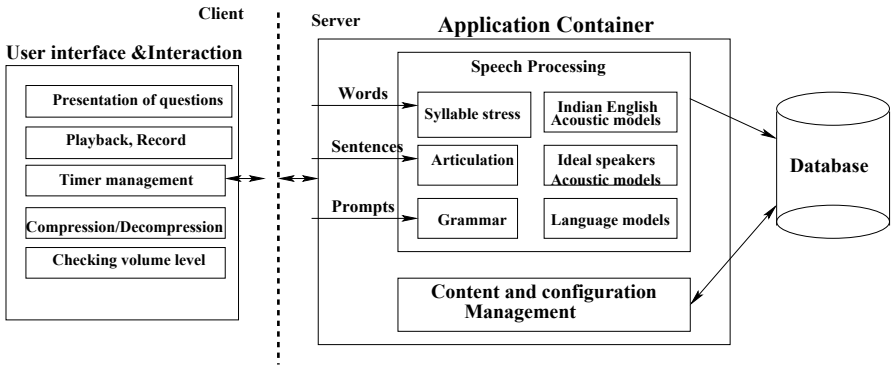

Fig. 1. Sensei System Architecture Overview

Sensei has three main components, *viz.*, user interaction and interface, speech processing and content & configuration management for various modules in the test.

2.1. User Interface and Interaction

User interface and interaction component renders an assessment test on the web browser using the content and configuration parameters for various modules in the test. This consists of displaying sentences and words to be recorded for evaluating articulation and syllable stress respectively, prompts to be played back to the candidate for grammar evaluation, recording user's utterances, compression/decompression of audio files to be sent over the network, enforcing a limited time for each of the modules, etc. This component also involves some amount of speech processing to make sure that the candidate's speech is recorded at proper volume level and warning the candidate in case no speech or only noise is recorded.

2.2. Speech Processing

Speech processing component, which resides on the server, uses the speech recognition engine to get the phonetic alignments, the confidence scores, the recognized utterance in case of grammar evaluation and such information. It supplies the grammar for recognition and specifies the acoustic models to be used for a specific module to get the required information from the recognition engine. General Indian English acoustic models are used to evaluate grammar and syllable stress and a small customized acoustic model, trained on the call center trainers (called ideal speakers from here onwards), is used to evaluate articulation. General Indian English acoustic model is chosen to ensure high recognition accuracy for grammar and better phonetic (and later syllabic) alignment in case of syllable stress. Speech processing component computes scores for each individual module and a combined score for overall assessment of the candidate. The score computation for various modules is discussed in detail in Section 3.

2.3. Content & Configuration Management

This component is to specify the content to be used for various modules and to control the nature and the difficulty level of the assessment test. An administrator can use this component to specify the words and sentences to be recorded by the candidate for evaluation of syllable stress and articulation respectively. The administrator can specify the audio files to be used for grammar evaluation. New audio files for grammar evaluation can be recorded and added to the already existing pool of questions. Another important functionality of this component is to change various configuration parameters, such as, maximum number of allowed attempts to record an answer, time allowed to complete a particular module, maximum number of times the candidate can listen to a grammar question, weights of individual parameter scores in the overall score and so on.

3. EVALUATION OF VARIOUS SPOKEN ENGLISH PARAMETERS IN SENSEI

The technical approaches used in Sensei to evaluate various spoken English parameters are now described.

3.1. Syllable Stress Evaluation

Syllable stress plays an important role in efficient spoken communication in English. The meaning of a word can change considerably based on the stress pattern of the constituent syllables. In this section, we briefly discuss our work [8, 9] on classifying English words spoken by Indian speakers into correct and incorrect classes based on the stress pattern of the constituent syllables.

Previous studies [10, 11] in this area have shown that stress is largely manifested through three basic prosodic features: fundamental frequency, duration and energy. Stressed syllables often exhibit higher values of these features as compared to their unstressed counterparts. In this work, we present two different methods for syllable stress evaluation: (a) two-class classifier and (b) single-class classifier model. Both the methods use eight prosodic features (three basic features and five derived features) computed at the syllable level. The feature computation process is explained in the following section.

3.1.1. Feature Computation

Each of the word utterances is time aligned with its phonetic spelling using the acoustic models and the pronunciation dictionary of a speech recognition system. A phone-to-syllable mapping for the word is then applied to get the syllable boundaries. The eight syllable level prosodic features are: (f1) Average fundamental frequency (F0), (f2) average energy, (f3) duration, (f4) average filtered energy, (f5) average energy X duration, (f6) F0 X duration, (f7) F0 ratio and (f8) energy ratio.

Fundamental frequency is estimated using a high resolution pitch estimation algorithm proposed in [12]. Features (f1)-(f3) are normalized with the corresponding values over the entire utterance to remove any speaker dependent variations. Feature (f4) uses a high-pass Butterworth filter with a cutoff frequency of 4 kHz to capture the energy content of the high frequency region . Feature (f7) is the ratio of the average fundamental frequency of the next syllable and that of the current syllable. Similarly, feature (f8) is the ratio of the average energy of the next syllable and that of the current syllable. These two features capture the temporal variations of the basic features across the syllables.

The two different methods for syllable stress evaluation are presented below:

3.1.2. Two-class Classifier

For each of the words in the system, correctly and incorrectly stressed models are trained using the corresponding spoken utterances. Instances of correctly stressed words are obtained from model speakers while the utterances from the agent speakers might or might not be correctly stressed. Human labelers labeled each utterance as either correct or incorrect based on the stress pattern of the word. The total number of features for a N-syllable word is $(8 * N - 2)$ since feature 7 and 8 do not exist for the last syllable. All the features are concatenated to form a single combined feature vector which is used to train a particular classifier. The classifiers used in this study are: Naive Bayes (NB), Decision Tree (DT), k-Nearest Neighborhood (KNN) and Support Vector Machine (SVM). Classification results corresponding to each of these classifiers are described in Section 4.3. Since this is a two-class model the training phase needs a set of incorrectly stressed words in addition to their correctly stressed counterparts.

3.1.3. Single-class Classifier

We have developed a classification technique by using only the correctly stressed utterances without using any incorrect utterance of the word [9]. This is motivated by the fact that for a given word the samples of the correct utterances are likely to form a compact region in the multi-dimensional feature space as there is only one way to pronounce the word with correct syllable stress pattern. The density of the correct class for a given word is estimated from the correctly stressed samples using the non-parametric Parzen window estimate with Gaussian kernels [13]. The conditional density of any point x in the multi-dimensional space belonging to the correct class C_p is given by

$$p(x|C_p) = \sum_{x_i \in C_p} K(x, x_i) \tag{1}$$

where x_i are the training samples, $K(x, x_i)$ is a Gaussian kernel with either Euclidean Distance (ED or with Mahalanobis Distance (MD). If x is a test utterance then it is assigned to the correct class if $p(x|C_p)$ is greater than a certain threshold θ. The threshold θ is selected based on the nearest neighbors of the test utterance.

$$\theta = min_{x_i \in C_p; x_i \in N_k(x)} p(x_i|C_p) + b \tag{2}$$

where $N_k(x)$ is the set of k nearest neighbors of the test utterance, x, from the correct class and b is a bias trained using a modified version of leave-one-out technique. More details about this approach can be found in [9].

3.2. Articulation Evaluation

Proper articulation of sounds plays an important role towards an effective communication between the call center agent and the customer. This achieves even higher significance in case of offshore call centers when the agent and the customer belong to different geographical regions and hence may have very different language accent. For example, it is often observed that Indian agents tend to interchangeably use sounds such as, /v/ and /w/, /z/ and /zh/ and so on. It is often required in an offshore call center that the agents should speak in a neutral accent (which is globally understood) without having to mimic a particular accent.

Researchers have tried various approaches in the past to evaluate articulation [14, 2]. In Sensei, the agent or the candidate is asked to record a set of sentences which are selected keeping in mind the common errors observed in articulation for various sounds. These sentences are also recorded by a set of ideal speakers[1] to train a customized acoustic model which is used as a benchmark for articulation evaluation. We generate a forced Viterbi alignment at the phone level for each of the sentences recorded by the candidate using those customized acoustic models. Phonetic confidence scores generated during the Viterbi alignment are then used to compute a score for the candidate as described below. Consider a phone p and the corresponding observation vectors aligned to it during the forced Viterbi alignment, $O_t, t \in \{b_p, .., e_p\}$, where b_p is the index of the first aligned frame and e_p is the index of the last aligned frame. Then, the confidence of the phone p, C_p, is computed as follows,

$$C_p = \frac{\sum_{t=b_p}^{e_p} \log(P(O_t|s_t^*))}{\sum_{t=b_p}^{e_p} \max_{1 \le j \le J} \log(P(O_t|s_j))} \tag{3}$$

where $P(O_t|s_t)$ is the rank likelihood [15] of O_t given HMM state s_t, $S = \{s_1, s_2, ..., s_J\}$ is the set of all HMM states and s_t^*, $t \in \{b_p, .., e_p\}$ is the Viterbi state alignment of the phone p while emitting the observation vectors $\{O_{b_p}, ..., O_{e_p}\}$.

The individual phone scores are aggregated to compute a combined score for the candidate. Let us denote the set of phones present in all the utterances recorded by the candidate as $\{p_1, p_2, ..., p_M\}$. Then the articulation score of the candidate is computed as follows:

$$S = \frac{\sum_{i=1}^{M} d_{p_i} C_{p_i}}{\sum_{i=1}^{M} d_{p_i}} \tag{4}$$

where, C_{p_i} is estimated from (3), and $d_{p_i} = e_{p_i} - b_{p_i} + 1$ is the number of frames aligned to phone p_i. The score S is later scaled to have the same dynamic range as that of the human ratings for articulation.

3.3. Spoken Grammar Evaluation

Ability to speak grammatically correct sentences is an important requirement for candidates for the call centers. An ideal way to evaluate a candidate's spoken grammar skills would be to extract grammatical errors from the candidate's free speech conversation with a machine. However, word error rates (WER) of current speech recognition systems for spontaneous speech makes it difficult. Moreover, language model which plays an important role in the recognition systems, makes grammatically incorrect sentences the least probable candidates to be decoded as output.

[1] Ideal speakers are the trainers at the offshore call center whom we assume to be using a canonical phonetic set in their pronunciation and have a universally understood accent.

Keeping in mind these constraints, we use a novel approach to evaluate the spoken grammar of a candidate as follows. The candidate listens to sentences spoken by Sensei and he/she is supposed to spot out grammatical errors in them and record the corresponding grammatically correct sentences. Based on the candidate's ability to spot the grammatical error in those sentences, candidate's spoken grammar skill is evaluated. Evaluation of correctness of grammar for each spoken sentence is depicted in Figure 2. Let us assume that the sentence spoken out by Sensei is called question, Q, and the corresponding sentence recorded by the candidate is called response, R. We want to map $R \rightarrow 1$ or $R \rightarrow 0$, depending upon whether the candidate was able to spot and correct the error or not respectively. The response speech from the candidate is decoded with Indian English speech recognition system, as will be explained in Section 4.2, to find an estimate of the response, $\hat{R}$. The speech grammar, A_i, used during recognition is restricted to a preselected set of sentences for the specific question asked. This restriction is because:

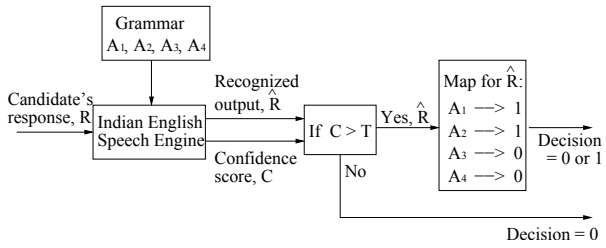

Fig. 2. Block diagram of grammar evaluation in Sensei.

1. for a given question the candidate is expected to spot a particular type of grammatical error, and
2. the candidate should not go out of the context of the question asked[2].

Along with the recognized output, $\hat{R}$, we also use the confidence score C from the recognizer to prevent any out of context response getting mapped to the sentences in the speech grammar. This helps reduce the false acceptance rate in the overall results for grammar evaluation. Confidence score for a response is computed using (4). Confidence score is used to perform first level decision as follows:

$$\text{decision} = \begin{cases} 0 \text{ if } C < T \\ \text{Map}(\hat{R}) \text{ otherwise} \end{cases}$$

where T is a threshold for confidence score and Map($\hat{R}$) is predefined mapping list, $A_i \rightarrow 1$ or 0. for pre-selected sentences in the grammar, A_i.

Confidence score is expected to become low if the candidate is speaking a sentence that is not present in the pre-selected set for the particular question asked. This would include the cases where candidate is giving answers that are completely out of context and the cases where candidate is not focusing on the grammatical error present. The threshold T was chosen to produce an acceptable value of false acceptance and false rejection rates.

[2]For example, given a question, *I own an big car*, candidate may possibly come up with one of the following answers:

1. *I own a big car*,
2. *I owned an big car*,
3. *I am owning a big car*,
4. *I own a big cat*, and
5. *I own an big cat*.

Out of these 1st, 3rd, and 4th sentences are grammatically correct. 2nd and 5th are grammatically incorrect. 1st is a valid correct answer. 4th and 5th are completely irrelevant to the context of the question asked. Although 3rd is in the context of the question asked, here the candidate is not focusing on the grammatical mistake in question. Ideally we would like Sensei to map 1st answer to 1, and 2nd, 3rd, 4th, and 5th answers to 0.

4. EXPERIMENTS AND RESULTS

4.1. Database

The database used to evaluate the performance of Sensei assessment consists of speech utterances collected during real-life assessment conducted on 243 candidates. Each of the candidates responded to 20 prompts(words/sentences/audio) each for syllable stress, articulation and grammar. Prompts were chosen randomly from a larger bank consisting of 100 syllable stress words, 200 articulation sentences and 204 grammar prompts. Each response (recorded articulation sentence, syllable stress word or grammar answer) given by the candidate was assessed by 3 independent human assessors to provide human ratings. The human assessor rating for syllable stress and grammar are either 1 or 0 depending upon whether the response is correct or not. For articulation, the rating vary from 1 to 4 where 4 corresponds to highest quality of articulation. The grammar questions were designed to test various spoken grammar parameters such as, propositions, articles, subject-verb agreement, word-order, tenses and so on.

Part of 204 prompts used in grammar evaluation were selected from a bigger set in a data-driven manner to maximize discrimination between good and bad candidates. Taking an example prompt *he are a boy*, it may be easy for anyone with little English knowledge to spot and correct grammatical error in it, resulting in less discrimination between good and bad candidates. Where as the prompt *he own a big car* may be able to discriminate between good and bad candidates well. An independent database has been used to select the discriminating subset of prompts, which has answers from candidates for grammar prompts from bigger set as well as human assessments as to whether the candidate is good or bad. Now, taking a particular prompt, q, let a be the number of candidates who have answered that prompt correctly and have also been rated as good candidates by the human assessors, b be the number of candidates answered wrongly and rated bad, c be the number of candidates answered correctly and rated bad, d be the number of candidates answered wrongly and rated good, then the prompt's discrimination capacity is computed using the equation given below:

$$S_q = \frac{a+b-c-d}{a+b+c+d} \tag{5}$$

Using values S_q for all the prompts considered top prompts were chosen.

4.2. Recognition Engine

For syllable stress and spoken grammar evaluation we used general Indian English recognition engine while for articulation evaluation we have used customized acoustic models trained on ideal speakers. For both the recognition engines, we trained context-dependent HMMs with context length 5 using 24 dimensional MFCCs, which are Linear Discriminant Analysis (LDA) transformed to 60 dimensions from 9 consecutive frames. Indian English acoustic models were trained on more than 500 speakers resulting in 130 hours of speech data. For articulation evaluation, we trained an acoustic model

on 70 ideal speakers. Each of the ideal speakers recorded 200 sentences, designed to cover a large phonetic diversity, resulting in 14000 utterances, approximately 22 hours of speech data.

4.3. Syllable Stress Experiments

The two-class method with four different classifiers (NB, DT, KNN and SVM) was trained for 13 words in the system. Each word had 30 utterances from model speakers (with correct stress pattern) and 227 utterances from agent speakers (mix of correct and incorrect stress patterns). The single-class models with two different distance metrics (ED and MD) were trained using only the utterances from the model speakers. Performance of the two-class and the single-class models were evaluated using an experimental speech database consisting of the same 13 words spoken by 75 agents (different from training), resulting in 975 utterances. The test utterances were labeled by two human assessors as either correct or incorrect based on the stress pattern of the constituent syllables of the individual words. Table 1 shows the average 3-fold cross validation performance of the different classifiers. Note that the two-class classifiers perform better than the single-class. However, single-class classifiers are important as they make it easier to add new words into the system with only correctly stressed utterances which are easier to obtain. The present deployed version of Sensei uses the DT-based two-class classifier resulting in a classification rate of 76.9% on a larger (93) set of words.

Table 1. Classification accuracy (in percent) for two-class and single-class syllable stress classifiers.

	Two-class				Single-class	
	NB	DT	KNN	SVM	MD	ED
exp-data	92.80	93.66	94.60	94.97	80.37	78.62

4.4. Articulation Experiments

Three human assessors listened to articulation recordings and independently assigned a score to each of the recording which was then normalized to be in the range $(0, 1)$. The ASR engine assigned a confidence score to each of the utterances using (4). In Table 2, we list the correlation coefficients between the machine confidence score and the human assessor scores. In Table 3, we list the correlation coefficients between the human assessors. Note that even human to human correlation is far from 1.0 which reflects the amount of subjectivity present in the human judgement. It was found that all the three human assessors provided the same rating only in 36% of the cases. However, in 92% of the cases their rating were either same or adjacent (1/2/3/4). This is the main reason why the correlation between Sensei score and average assessor score is higher than the correlation with any of the individual assessors.

Table 2. Correlation coefficients between Sensei scores and human assessors scores.

Assessor 1 vs Sensei	0.54
Assessor 2 vs Sensei	0.51
Assessor 3 vs Sensei	0.52
Average assessor score vs Sensei	0.56

Table 3. Inter-human correlation coefficients.

Assessor 1 vs Assessor 2	0.75
Assessor 2 vs Assessor 3	0.63
Assessor 3 vs Assessor 1	0.80

4.5. Spoken Grammar Experiments

Results of automatic grammar assessments on the database mentioned above are compared with the corresponding human scores to find out overall accuracy. Table 4 summarizes the corresponding results. The first, second, and third rows show accuracy when Sensei assessments are compared with the assessments of first, second, and third human assessor respectively. In some cases, mainly in deciding whether to accept an answer as valid or not, the ratings from human assessors may not match and hence all human assessors agree in 87% of the cases. Fourth row in the table shows accuracy only on this subset of the data. The accuracy numbers achieved are highly encouraging considering the small acoustic difference among the expected responses from the candidate for a question. For example, considering sentences, *He own an big car*, *He owns an big car* and *He owns a big car*, the acoustic difference among these sentences is very small.

Table 4. Accuracy of Sensei grammar scores compared with the human scores.

Sensei vs Assessor 1	78.4%
Sensei vs Assessor 2	77.8%
Sensei vs Assessor 3	78.6%
Sensei vs Assessors (on 87% of data where all the human assessors agree)	81.1%

In order to provide a combined assessment score to a candidate, the scores from individual parameters are linearly combined using weights decided by the required profile for the call center. For the deployment, weights of 0.3, 0.2 and 0.5 were chosen for articulation, syllable stress and grammar respectively. An overall correlation of **0.8** was observed between the overall Sensei scores and human scores on the test database described earlier. On the same dataset inter-human correlation was found to be 0.91.

5. CONCLUSION

In this paper, we described a system, called Sensei, for evaluation of spoken English skills of a person on multiple parameters. We described how different approaches have been used in Sensei to evaluate various spoken English parameters, *viz.*, spoken grammar, articulation and syllable stress. It was also shown that the scores generated by Sensei are close to the corresponding human scores. Since there is a considerable amount of subjectivity even among different human assessors, Sensei accuracy with a human assessor should be compared with the corresponding human-to-human accuracy. Sensei can benefit the offshore call center industry in a significant manner as it brings scalability, objectivity and cost-effectiveness into the assessment process.

6. DISCUSSION & FUTURE WORK

Note that the errors committed by the recognizer in the phone-level time alignment are propagated to the syllable boundaries which affect the syllable stress evaluation. Work is in progress to develop algorithms that will automatically refine the syllable boundaries to counter these errors. For grammar evaluation, since the possible responses to a given question are acoustically close to each other, sometimes a response is mis-recognized as another response. This can possibly be avoided by using a higher weight in the matching for the regions where the responses are different.

7. ACKNOWLEDGEMENT

The authors would like to deeply acknowledge Lyndon D Silva, Mridula Bhandari, Rajni Bajaj, Leon Dawson, Viren Singh Ghuman and Tanmay Roy from IBM Daksh for providing speech training data, human labels and content during development of Sensei. The authors would also like to acknowledge Vivek Tyagi from IBM India Research Lab for his valuable feedback and suggestions.

8. REFERENCES

[1] J. Tepeerman, J. silva, A. Sethy, and S. Narayanan, "Robust recognition and assessment of non-native speech variability," in *In Proc. of International Conference on Intelligent Systems And Computing: Theory And Applications*, Ayia Napa Cyprus, July 2006.

[2] K. You, H. Chang, J. Lee, and Wonyong Sung, "A handheld english pronunciation evaluation device," in *International Conference on Consumer Electronics*, January 2005.

[3] K. Jenkin and M. Scordilis, "Development and comparison of three syllable stress classifiers," in *in Proc. of Intl conference on speech and language processing*, Philadelphia, October 1996.

[4] L. Oppelstrup, M. Blomberg, and D. Elenius, "Scoring children's foreign language pronunciation," in *in Proc. of FONETIK*, Goteborg, May 2005.

[5] E. Fosler-Lussier, "Contextual word and syllable pronunciation models," in *International Workshop on Automatic Speech Recognition and Understanding*, Keystone Colorado, December 1999.

[6] Ordinate, "www.ordinate.com," .

[7] GlobalEnglish, "www.globalenglish.com," .

[8] A. Verma, K. Lal, Y Y Lo, and Jayanta Basak, "Word independent model for syllable stress evaluation," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, May 2006.

[9] A. Parate, A. Verma, and Jayanta Basak, "Evaluation of syllable stress using single class classifier," in *Proc. Interspeech*, Antwerp, August 2007.

[10] Rosaria Silipo and Steven Greeberg, "Automatic transcription of prosodic stress for spontaneous english discourse," in *Proc. International Conference on Phonetics*, Apr. 1999, pp. 2351–2354.

[11] J. Tepeerman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[12] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 39, pp. 40–48, Jan 1991.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd edition, New Yorkk; Wiley, 2001.

[14] H. Franco, L.Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. of ICASSP*, 1997.

[15] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Robust methods for using context-dependent features and models in a continuous speech recognizer," in *Proc. of ICASSP*, April 1994.