Development of the 2012 SJTU HVR System

Hainan Xu
Shanghai Jiao Tong University
800 Dongchuan RD. Minhang
Shanghai, China
xhnwww@sjtu.edu.cn

Yuchen Fan
Shanghai Jiao Tong University
800 Dongchuan RD. Minhang
Shanghai, China
fyc0624@sjtu.edu.cn

Kai Yu
Shanghai Jiao Tong University
800 Dongchuan RD. Minhang
Shanghai, China
kai.yu@sjtu.edu.cn

ABSTRACT

Haptic voice recognition (HVR) is a multi-modal text entry method for smart mobile devices. It employs haptic events generated by speakers during speaking to achieve better efficiency and robustness for automatic speech recognition. This paper describes the detailed design of the 2012 SJTU submission for the HVR Grand Challenge. During the design, a new perplexity metric using conditional entropy is proposed to evaluate the potential search space reduction of a haptic event without speech input. A number of new haptic events are evaluated both theoretically and experimentally in detail. The final submission system uses the haptic event of initial letter plus final letter and reduces word error rate by 76% compared to the baseline initial letter event.

Categories and Subject Descriptors

I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing—Speech recognition and synthesis

Keywords

haptic voice recognition; multi-modal interface

1. INTRODUCTION

Mobile devices nowadays play an important role in people's lives, yet text entry methods on mobile devices are never satisfying. The most common way to enter text on mobile devices is by tapping on the QWERTY keyboard on the touch screen, and for its lack of haptic feedback, it is highly error prone and slow. As speech is the most natural way of communication of human beings, a great deal of efforts have been made by experts to make automatic speech recognition (**ASR**) possible on mobile devices. However, due to their limited computing resources, large vocabulary continuous speech recognition is nearly impossible to accomplish. What's more, mobile devices are commonly used in noisy environments. No matter what recognition architecture is used, noise will always significantly degrade the recognition performance. Two major solutions have been used to address these problems: First is distributed speech recognition, where the waveform or extracted acoustic vectors are transmitted over the network to the servers where recognition is actually performed and results are sent back. An example of this is Google's speech recognition service. The second solution is that the recognizer predefines the grammar and limit the vocabulary of recognition task so that the search space is reduced, one of the examples is the voice control on iPod touch 4, where the grammar is fixed to be among "play album" followed by an album name, "play songs by" followed by a singer's name, "next song", etc.

Haptic voice recognition (**HVR**)[1] aims to solve the problem by introducing additional information. In HVR, mobile users provide additional haptic inputs on touch screens while they are speaking to the device. Based on different designs of the haptic events, this information could be used by the recognizer in a number of ways to help the recognition, such as reducing search space, helping acoustic model adaptation etc. Among them, search space reduction is the primary advantage of HVR. With a reduced search space, an HVR system requires less computing resource than a standard ASR system does; furthermore, haptic events could potentially compensate for lost or distorted acoustic information due to effect of the noise, making the system more robust in noisy environments. Hence, in this paper, search space reduction is the focus during the design of haptic events.

What a haptic event does to speech recognition is not all positive though. Firstly, the fact that users need to provide haptic inputs to the recognizer while speaking inevitably slows down the input speed, so a trade-off between recognition accuracy and input speed must be made. Secondly, the haptic input provided by the user may contain errors, and thus the system needs to either be less error prone or be able to handle erroneous inputs. Otherwise the erroneous haptic event could prevent the right word or word sequence to be recognized.

This paper describes the design of SJTU submission for the task 1 of the HVR Grand Challenge 2012[2]. The objective of task 1 of the challenge is to design innovative haptic events for HVR and methods for generating these events using touch-screen inputs. This task also imposes a time limit of 80 words per minute (**WPM**) on the overall input. The challenge release contains a baseline HVR system which uses the initial letter sequence of user utterances as the haptic event. Experiments show that the use of haptic events significantly reduces the word error rate (**WER**). In this paper, a number of more informative haptic events

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.
Copyright 2012 ACM 978-1-4503-1467-1/12/10 ...$15.00.

are proposed and achieve further gains. The final submission system obtains 76% relative gain over the baseline HVR system.

The rest of the paper is organized as follows: In section 2, a new metric is proposed to evaluate the effectiveness of a haptic event without speech input. Section 3 describes the design and implementation of the SJTU HVR systems. Experiments are detailed in section 4, followed by the conclusion and future work.

2. EVALUATION OF SEARCH SPACE REDUCTION IN HVR

As indicated in the previous section, the primary advantage of incorporating haptic events into speech recognition process is to reduce the search space so that the search complexity can be reduced and the accuracy can be increased. Although this effect can be evaluated by WER from actual HVR experiments, it is not convenient to do so. To design useful haptic events, it is important to evaluate how much a haptic event can help to predict words without speech input. The evaluation can be done by investigating the effect of haptic events on language model. This section will investigate appropriate metric for the purpose. To simplify the analysis and be consistent with the HVR challenge setup, the language model considered here is just *word loop*, i.e. unigram language model with uniform distribution over all words of the recognition vocabulary. Similar idea can be extended to N-gram language model.

In conventional speech recognition, the prediction power of a language model is measured by perplexity[3]. For a recognition task with vocabulary V, the *perplexity* (**PPL**) of a word-loop language model is defined as

$$PPL = 2^H$$

where H is the entropy of an unigram language model, defined as

$$H = -\sum_{w \in V} P(w) \log_2 P(w)$$

In the case of word-loop, given the number of words of the vocabulary V, denoted as $|V|$, $P(w) = \frac{1}{|V|}$ is a uniform distribution. Hence, the entropy becomes

$$H = \frac{1}{|V|} \sum_{w \in V} \log_2 |V| = \log_2 |V| \tag{1}$$

and consequently the perplexity $PPL = |V|$. From the definition, perplexity is the complexity of the language model, hence, the smaller it is, the more prediction power the language model has.

In HVR, the search space is not only defined by the language model, but also constrained by observed haptic events. Considering the case of offline batch processing, given a haptic event associated with a word, the prior probability of the word in a word-loop language model becomes

$$P(w|e) = \begin{cases} |V_e|^{-1} & w \in V_e \\ 0 & w \notin V_e \end{cases}$$

where V_e denotes the vocabulary associated with haptic event e. Hence, the entropy of the word-loop language model conditioned on the haptic event e is defined as

$$H(e) = -\sum_{w \in V} P(w|e) \log_2 P(w|e) = \log_2 |V_e| \tag{2}$$

The overall entropy of the conditioned language model can then be represented by the expectation of the conditional entropy. Considering that the prior probability of the haptic event is also a uniform distribution, i.e. $P(e) = \frac{|V_e|}{|V|}$, the entropy of haptic event conditioned word-loop language model is defined as

$$H_{hap} = \sum_{e} P(e) H(e) = \frac{1}{|V|} \sum_{e, w \in V_e} \log_2 |V_e| \tag{3}$$

Since V_e is a subset of V, $|V_e| \leq |V|$. By comparing equation (3) to equation (1), it is then easy to find that H_{hap} is always no greater that H, consequently the perplexity $PPL_{hap} = 2^{H_{hap}}$ is no greater than PPL. This demonstrates that incorporating accurate haptic events will almost always help to reduce the perplexity, i.e., increase the prediction power of the corresponding language model.

In addition to the theoretical estimation of perplexity as described before, in practice, it is also useful to estimate how well a particular language model matches the data of a specific test-set. As an analogy of the conventional definition of the perplexity on test-set data, the experimental perplexity for HVR is defined as $PPL_{hap} = 2^{LP_{hap}}$, where LP_{hap} is the log probability of the test data conditioned on haptic events. In the word-loop language model case, it is defined as

$$LP = -\frac{1}{K} \sum_{k=1}^{K} \log_2 \frac{1}{|V_{e(k)}|} \tag{4}$$

which $e(k)$ denotes the haptic event of the k^{th} word, and K is the total number of words in the test-set.

In this paper, both theoretical and experimental perplexity of the conditioned word-loop language model will be used to evaluate the search space reduction.

3. DESIGN AND IMPLEMENTATION OF SJTU HAPTIC EVENTS

During the design of SJTU haptic events, two assumptions are made from the understanding of the HVR challenge rules. Firstly, haptic events are incorporated into speech recognition in an offline mode, i.e. actual speech recognition is performed after receiving all haptic events and speech data. Secondly, word-loop language model is used. With these assumptions, the effect of a haptic event is simply to change the structure of the word network for decoding.

As indicated in section 1, three important factors should be considered when designing a haptic event for HVR:

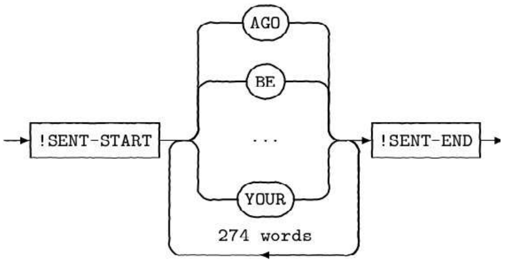
1. Search space reduction
2. Input reliability
3. Input speed

Previously, *initial letter* has been used as a typical haptic event for HVR, referred to as **IL**. This is also the baseline system provided by the HVR organization committee. In this paper, in addition to **IL**, three haptic events are proposed. The advantage, disadvantage and implementation of each haptic event are discussed in the below sub-sections.

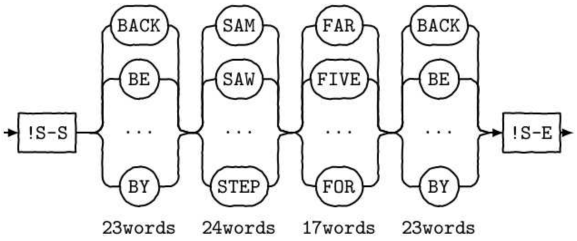
3.1 Initial Letter (IL)

In the baseline system, users need to input the initial letter of each word while speaking. With this haptic inputs, the search space for each word shrinks and only contains the words having the same initial letters.

Taking the development dataset of the HVR challenge as an example, there are altogether 274 words in the vocabulary. Suppose a user wants to input "BUY SOME FRUITS BACK", with a word-loop language model, the search space is formed by the whole vocabulary, as depicted in figure 3.1. The PPL of the word net is 274.



With IL event, the word net changes to:



From equation (3), the PPL reduces from 274 to 14. This will result in a significant reduction in search space. What's more, the initial letter information also implies the number of words in the sentence, which helps to reduce the insertion/deletion error as well.

As users only need to input one letter per word, and initial letter of a word is usually directly related to its pronunciation, IL has a pretty good input speed and users are not likely to make input errors. Details of input speed and input reliability will be shown in section 4.

The disadvantage of IL is obvious: IL is far from being sufficiently informative as the recognizer confuses a lot between words with the same initial letter, like "ford" with "for", "like" with "light".

3.2 Initial Letter plus Length

As discussed in the previous section, the baseline HVR system confuses between words with the same initial letter as well. To further reduce the PPL, additional information needs to be added. In this section, word length is investigated as the additional information.

Theoretical analysis shows that, by restricting initial letter and word length information, the PPL of the recognition network is down to 2.78. And also, as longer words generally have longer pronunciations and vice versa, by constricting the length of a word, the recognizer also has implicit knowledge of how long the word is in the recording, and this implicitly gives the recognizer the "boundary of word" information.

Ideally, the accurate length of word information would be most effective in reducing PPL. This is referred to as *Initial Letter plus Accurate Length* (**ILAN**). Although ILAN yields much better PPL than IL, yet a practical problem exists in the that approach: users may not be able to remember the length of a word immediately while speaking, especially for long words, and this would result in more input error and/or slow down the user input. Therefore, it is necessary make compromises.

A series of experiments have been performed to investigate the reliability of the word length event. It is found that in most cases people can remember the exact length of a word when the length is less than 5. Under that assumption, a new event, *Initial Letter plus Fuzzy Length* (**ILFN**) is proposed. Theoretical analysis shows that the PPL of the recognition network is 6.32 on the development data, better than IL. Although it is not as good as ILAN is, this design made this haptic event possible to implement, details about which will be in the following paragraphs.

Once the haptic event is determined, a friendly interface needs to be implemented to efficiently generate it. For most modern mobile devices, the touch screen could handle two types of input: taps and swipes. When these inputs are performed on a QWERTY keyboard, the coordinates of the tapping could be mapped to the English alphabet and the distance of the swiping could represent word length information. Those two types of input are sufficient for us to implement the haptic event introduced above. Fig 1 illustrates how haptic events are interpreted on the QWERTY keyboard.

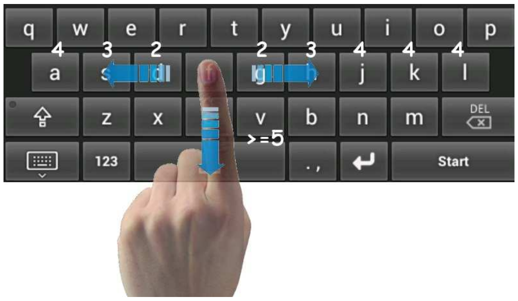


Figure 1: User Interface for ILFN

For simplicity, all user input on the touch-screen are interpreted as swipes, and the distance of a swipe is defined as the actual distance on touch-screen divided by the width of a key (a quick tap-and-release would be viewed as a swipe with distance 0). The tapping point on the keyboard is recorded as the initial letter of the event. And there are 3 legal swiping directions: left, right and down. For a left or right swipe, the distance plus one denotes the length that the user wants to input; for down swipes, it denotes that the word length is greater or equal to 5. For example, in the picture, if the user wants to input the word "fun", he will tap on the Key F and swipe to the right or left with distance 2, which means he should either release on the Key S or the Key H; for the word "fill", he will tap on the key F and then release on the Key A or J; for the word "family", the user need to tap on F and release on the Key C or the Space. For single-letter words like "I" and "A", the user need to simply tap on the corresponding Key and release on the same key.

Although it seems complicated from the description, in practice, this input method is highly efficient and easy to learn.

3.3 Initial Letter plus Final Letter (ILFL)

Adding the word length information is not the only solution. Theoretical analysis results show that, restricting both initial letter and final letter further reduces the PPL to 1.99, better than the PPL of ILAN (2.78).

A disadvantage of ILFL is that, unlike the initial letter information of a word which is easy for people to remember, the final letter is relatively harder for people to come up while speaking; another disadvantage is that it requires the user to search for 2 keys for each word in inputting, both of which slow down the input speed and reduce input reliability.

3.4 Initial Letter plus Final Phone (ILFP)

Given the effectiveness of the information about word end, it is useful to find a compromise way with both reliability and speed for ILFL. Unlike the final letter information which may be a little hard to remember immediately, the final phone of a word is easier to come up. However, it is impractical to include all final phones. Considering both the convenience for the users and their effect on reducing PPL, only 4 types of final phones are used:

1. m, n and ng

2. iy and ih

3. s, z, th and dh

4. l

This yields a theoretical PPL of 5.79. Although it is not as low as ILFL, as it does not require information that is hard to remember, combined with the fact that for each word, user only needs to search for one key on the keyboard, the input speed of this method is significant faster than ILFN and ILFL.

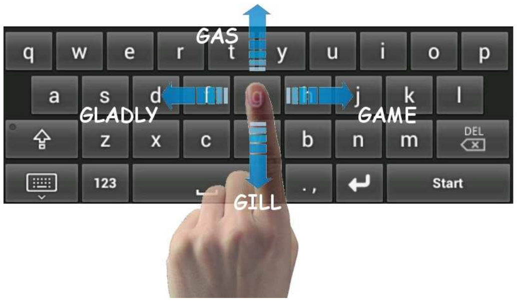


Figure 2: User Interface for ILFP

As shown in Fig 2, the initial letter information is determined by the tapping position on the keyboard. After the tapping, if the final phone of the current word falls in the 4 categories, the user could swipe in the corresponding direction so that this final phone information is added to the event. In any situation if the user is not sure of the final phone, he could always directly release without swiping and then only the initial letter information is recorded in the haptic event.

4. EXPERIMENTS

In this section, experiment results of different haptic events are presented. The ASR (speech recognition without any haptic event) results are also shown so that the effect of haptic events to speech recognition could be clearly seen. The organizer provided an acoustic model for speech recognition, and a recipe script to automatically adapt the acoustic models to different speakers on different noisy environments. These models and scripts are used in both development and evaluation.

The designed haptic events were evaluated on two data sets. The development data consists of both audio recordings and text provided by the organizer of the challenge. There are 80 sentences in the test-set and 243 in the training set, spoken by 6 speakers. 4 sets of data with different noise level were provided: clean, 20dB, 15dB and 10 dB.

As for evaluation, the organizer only provided text prompt. Audio data were recorded by 2 speakers in SJTU. There are 200 sentences in the test set and 40 sentences in the train set which is for adapting the models. Train-data and test-data are made separately. Test recordings are made while the speaker is also inputting the haptic events on the mobile phone, and the haptic events (which contain errors) created in that process are used in recognition.

4.1 Haptic Events Comparison

To choose the best system for the SJTU submission for the challenge, all haptic events were evaluated in terms of search space reduction, recognition performance, input reliability and speed.

4.1.1 Search Space Reduction

The analysis of search space reduction of all haptic events are listed in Table 1. The experimental PPL are calculated using the development and evaluation data with equation 4.

Table 1: Perplexity of Different Haptic Events

PPL	ASR	IL	ILAN	ILFN	ILFL	ILFP
The.		14.05	2.78	6.32	1.99	5.79
Dev.	274	13.18	2.38	4.76	2.03	4.01
Eva.		14.00	2.62	4.20	1.93	6.03

From table 1, all new haptic events obtained significant reduction in PPL. It is expected that the speech recognition performance will be similar.

4.1.2 Recognition Performance

As the development release does not contain data about the haptic events, the newly proposed haptic events were generated from the given transcription in development set (which means that those haptic events do not contain errors). The modified word networks associated with each haptic event were then used for decoding in various noise environments. The results on the artificial data are shown in Fig 3.

From Fig 3, all haptic events help in reducing the WER, and makes recognition less affected by noise. The ordering is also the same as in the PPL table 1.

In addition to the development, similar experiments were also done on the evaluation data. To simulate noisy environments, artificial noise were added to the audio recordings.

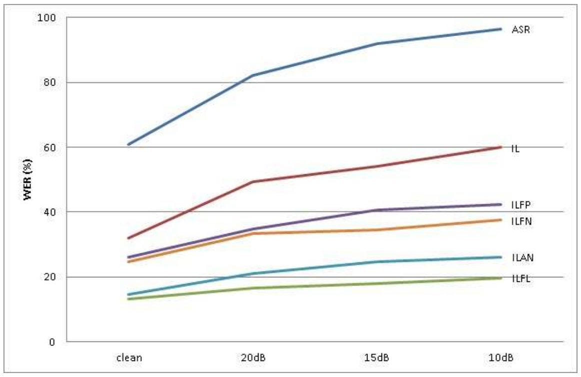


Figure 3: WER On Dev Set With Different Noise Level

The noise added was "babble" from the NOISEX-92 noise database[4], and FAnT tool[5] was used to control the data-noise-ratio. Test results on evaluation data are shown in Fig 4.

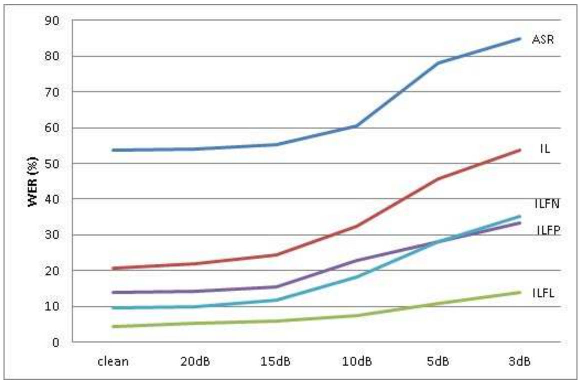


Figure 4: WER On Eval Set With Different Noise Level

The results on the evaluation data have exactly the same trend as the development data. This shows that the proposed haptic events have consistent effect on speech recognition.

4.1.3 *Input Speed and Reliability*

As discussed before, complex haptic events could slow down the user input. Experiments are conducted to see if our haptic events design satisfies the input speed limit imposed by the challenge.

Input reliability is another important factor. Here input reliability is measured by input error rate (**IER**). IER is defined as the number of errors made during the input divided by the number of total words entered. For instance, in the ILFL haptic event, if the user mistakenly input the word "play" as with initial letter O and final letter T, it will be counted as 2 mistakes.

The experiment is conducted as follows: the subjects first spent enough time in familiarizing themselves with the corresponding haptic event input by exercising random sentences on the mobile phone. After the subjects feel confident, they are asked to input all 200 test sentences on mobile phones during which they also speak the test sentences, and their input speed and input reliability is monitored. Table 2 is our experiment results on those 4 types of haptic events.

Table 2: Input Speed and IER of Different Haptic Events

	IL	ILFN	ILFL	ILFP
Input Speed (WPM)	103	85	83	89
IER (%)	0.63	2.59	1.26	0.91

It could be seen from the table that for their complexity, the other 3 haptic events require more time to input than the baseline haptic event. The slowest is ILFL, for the fact the final letter information is not as easy to remember as other information, and the users need to search for two keys when inputting each word. The IER for ILFL is twice than IL, which is consistent with the fact that ILFL requires 2 key input per word while IL requires 1; ILFP has an IER in between IL and ILFL, and is faster than ILFL. [1]

4.2 Performance of the Final SJTU HVR System

In the previous sections, the performance of different haptic events have been shown. It could be seen that the ILFL has the lowest WER in all settings; although it has the lowest input speed among others as well, it is still higher than the required 80WPM limit; its average IER is acceptable too, so the ILFL system is chosen as our final HVR system for submission.

Table 3: Summary of 2012 SJTU HVR System

	ASR	IL	ILFL
PPL	274	14.05	1.99
WER(% in dev,15dB)	91.89	54.22	18.07
WER(% in eval,15dB)	53.96	24.39	5.89
Input Speed(WPM)	-	103	83
IER(%)	-	0.63	1.26

From table 3, the submitted system has achieved about 76% relative WER reduction compared to the baseline IL system.

4.3 Further Discussion

4.3.1 *Recognition Speed*

Recognition speed is an important factor for evaluating a speech recognition system, especially when the system works on a resource-limited environment like a mobile phone. We conducted experiments testing how much time it takes the HVR system with different haptic event design to recognize the data provided in the development release. The data contains 80 sentences read by 6 speakers. The PC for testing is on Ubuntu 12.04 32-bit with a Intel Core i3-2120 (3.30GHz) dual processor and 2GB memory.

From the results shown in table 4, it is obvious to see that, by introducing haptic events, the recognition time is greatly reduced in all settings. Compared with standard ASR, the

[1] According to the design of ILFP, the user does not have to provide the final phone information even if the final phone does fall in the 4 categories. In the experiment result, 23 such case occured, while the total number of words whose final phones fall in the 4 categories is 581.

Table 4: Decoding Time (Second) of Different Haptic Events

SNR	ASR	IL	ILAN	ILFN	ILFL	ILFP
clean	523.02	11.92	3.34	5.60	3.17	4.81
20dB	748.90	22.14	4.77	9.36	4.24	7.74
15dB	761.41	25.10	5.37	10.66	4.75	9.01
10dB	759.51	28.74	5.80	11.82	5.14	10.02

ILFL reduced recognition time by a factor of more than 100, which means that the HVR system requires much less computing resources, making it more likely to be able to run on mobile devices.

4.3.2 *Effect of Vocabulary Size*

As so far, all tests have been on a small vocabulary set. Due to the reduction of PPL, haptic events help greatly in reducing WER in recognition. But whether haptic events could still have significant effect on large vocabulary sets remains to be seen. This section adopts the same PPL analysis introduced in section 2 and check whether the gain of haptic event will retain for larger data set.

We did PPL analysis on two larger vocabulary sets: DARPA Resource Management[6] and Wall Street Journal[7]. RM has a vocabulary size of 990 while WSJ has a vocabulary size of 20k. The PPL analysis gives the following results on them:

Table 5: Perplexity on Different Vocabulary Sets

	ASR	IL	ILAN	ILFN	ILFL	ILFP
HVR	274	14.05	2.78	6.32	1.99	5.79
RM	990	50.62	6.08	26.37	5.06	20.33
WSJ	20153	958	107	654	83	348

It could be seen that the effect on reducing PPL of haptic events are still significant in larger vocabulary sets and thus there are reasons to believe that haptic events could still have the significant effect in improving recognition performance in larger datasets. As language models are often used in large vocabulary datasets which also reduces PPL, it could work with haptic and altogether, they can achieve even better results in PPL reduction.

5. CONCLUSIONS

Haptic Voice Recognition (HVR) is useful to achieve highly efficient and robust speech recognition. In this paper, a PPL metric using conditional entropy is proposed to evaluate the effect of haptic events without speech input. The metric has shown consistent results with the final WER in experiments. A number of new haptic events are also introduced in this paper that helps improve speech recognition performance. In the final system we submitted, it reduces WER of recognition by 76% in 15dB recordings compared to the baseline haptic event, while still satisfying the 80WPM limit imposed on input speed.

The future work consists of several subtasks. Research on how to further improve the HVR model and implementation introduced in this paper will be conducted. This includes studies on incorporating haptic events with language models on large-vocabulary speech recognition tasks. We are also going to design an better user-interface to further increase input speed and reduce input error rate by utilizing gesture recognition methods or other technologies. Another aspect of our study is how haptic events could be used in other ways, like those not related to spelling information of words, and haptic events that are not solely aiming for reducing search spaces.

6. ACKNOWLEDGMENTS

The authors would like to thank AISpeech Ltd for providing support for the work.

7. REFERENCES

[1] Khe Chai Sim. Haptic voice recognition: Augmenting speech modality with touch events for efficient speech recognition. In Dilek Hakkani-TÃijr and Mari Ostendorf, editors, *SLT*, pages 73–78. IEEE, 2010.

[2] HVR Grand Challenge 2012. `http://speech.ddns.comp.nus.edu.sg/HVRGrandChallenge2012/index.php`.

[3] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-5(2):179 –190, march 1983.

[4] Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

[5] Filtering and Noise Adding Tool. `http://dnt.kr.hsnr.de/download.html`.

[6] P. Price, W. Fisher, J. Bernstein, and D. Pallet. The DARPA 1000-word resource management database for continuous speech recognition. In *IEEE ICASSP-88*, volume 1, pages 651–654, New York, 1988. IEEE.

[7] D. B. Paul and J. M. Baker. The design of the Wall Street Journal-based CSR corpus. *Proc. DARPA Speech and Natural Language Workshop*, 1992.