Silent Voice Elements for Text Input

Peng Teng and Yunde Jia
Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology
Beijing 100081, China
Email: {tengpeng, jiayunde}@bit.edu.cn

***Abstract*—Speech input systems could not work well in noisy environment, and their usage often makes a leakage of information. To avoid these problems, this paper proposes a concept of Silent Voice Elements, called sivels for short, and a novel articulators-operated text input method with sivels. Sivels are easy-to-recognized phonemes of soft whisper in their tissue-conducted vibration signals. The selection of sivels is a combinational optimization problem which is solved by using a heuristic search algorithm. Encoding text with sivels similarly to Morse code, one can input text accurately by speaking corresponding sivels. Experimental results demonstrate that our method selects a set of sivels with perfect recognizability, and the proposed sivel-based text input also gives an performance with sufficient efficiency.**

Keywords-text input; silent voice element; silent speech interface.

I. INTRODUCTION

Speech input is expected to become the principal text input method replacing keyboards. People have built and deployed numerous speech input systems for various applications. But in noisy environment, these systems have serious degradation in their performances, and can not work well. Besides, speech may be considered as unwanted noise, and often makes a leakage of information as well. Silent Speech Interface(SSI) [1] is a promising technology that can be employed to overcome these problems. For example, [2] captured articulatory movements using Electromagnetic Articulography (EMA) sensors and mapped them into phonemes; [3] aimed to recognize speech from data captured by Surface electromyography (sEMG) on articulatory muscles; [4] investigated an approach which directly recognizes "unspoken speech" in brain activity measured by Electroencephalographic (EEG) signals. Most of SSIs are still on the stage of laboratory research.

There is another SSI called Non-Audible Murmur (NAM) microphone [5], [6], a high-sensitivity contact microphone attached on the skin over the soft tissue in the orofacial region. [7] and [8] reported the NAM enhancement to audible speech for human-human communication. Compared with the sensor data acquired by other SSIs, NAM signal is a tissue-conducted acoustic signal which can provide a more direct and stable representation to the real speech and with insensitivity to noise. However, NAM recognition is difficult to be adopted as the underlying recognition technology for text input, because it could not deliver a low-error performance [9] owing to the poor quality of NAM signal in addition to the intrinsic difficulties for machines understanding human languages.

The goal of this paper is to present an alternative text coding scheme for developing an articulators-operated text input method with high accuracy as well as acoustic environment insensitivity. Specifically, we propose a concept of SIlent Voice ELement called sivel for short, and develop a novel method of articulators-operated text input where text is encoded with sivels. Sivels refer to easy-to-recognized phonemes in tissue-conducted signals of soft whisper, regardless of their linguistic meanings. Similar to Morse code, a user can encode text with sivels using a customized scheme and input text accurately by speaking corresponding sivels in soft whisper. The sivel-based text input method can work without the intervention of hands like speech input, and without noise sensitivity or information leakage. In addition, because of the customized scheme of text coding, sivel-based text input avoids the intrinsic difficulties for machines understanding human languages, and has the potential to provide an articulators-operated text input method for speech disorder people.

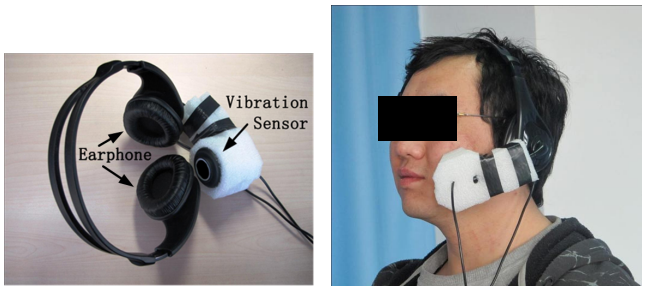
II. SILENT VOICE ELEMENTS (SIVELS)

In this section, we introduce the concept of sivel, and empirically select a set of sivels as an example. Then experiments on sivel recognition are preformed to evaluate the efficiency of these sivels as code elements, and the results will help the development of a general sivel selection method in the next section.

A. Tissue-conducted Soft Whisper

Soft whisper is a kind of low-amplitude sound that people pronounce without the vibration of vocal cords, and is not expected to be heard by others. When speaking soft whisper, one's articulators figure the vocal tract with certain shapes. Airflow out of the lung flows through the vocal tract and generates noise at its constricted segments. Soft whisper is namely the mixture of the noise and its vocal-tract resonance, and also a vibration of air. The vibration stimulates the vocal-tract wall, and some vibration energy transmits to the surface of one's head. With vibration sensors

 ISBN: 978-1-61208-195-3



(a) (b)

Figure 1. The headset of a experiment system(a) and its implementation(b).

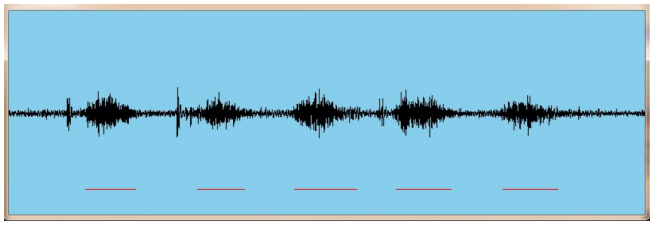


Figure 2. A vibration wave signal recorded by our experiment system where straight line segments denote the corresponding time when soft whisper is being pronounced.

placed on the skin of orofacial regions, the vibration can be detected and recorded, i.e., the tissue-conducted signal of soft whisper. We designed an experiment system as shown in Fig.1. An example of a vibration signal detected by our system is shown in Fig.2 where straight line segments denote the corresponding time when soft whisper is being pronounced.

B. An Example of a Sivel Set

Sivels are easy-to-recognized phonemes in tissue-conducted signals of soft whisper. Served as code elements for text coding, sivels are required to be stable and distinguishable in their signal patterns, so that they can be recognized easily and accurately.

We empirically select a set of sivels from whispered phonemes by considering their pronunciations. When different phonemes are pronounced in soft whisper, different articulatory positions make the vocal tract present different resonance characteristics. Since the source of soft whisper can be seen as the white noise, it can be assumed that signal pattern of a whispered phoneme is mainly resulted by articulatory position when one is pronouncing it, and phonemes with significant differences in their articulatory positions have significant differences in their signal patterns. Consequently, we select whispered phonemes /a/, /ə/, /i/ ,/ɒ/ ,/u/ (in English) from International Phonetic Alphabet (IPA) as sivels initially. Our reasons can be summarized as follows.

- These phonemes are all vowels, i.e., phonemes pronounced with an open vocal tract, so that their tissue-conducted signals can be detected and processed easily due to their relatively high amplitudes.
- These phonemes are all monophthongs, so the articulatory positions are almost unchanged during pronouncing; this makes their signal patterns stable.
- There are significant differences in their articulatory positions according to the IPA vowel diagram which shows the correlation of a monophthong and its corresponding articulatory position, so their signals could be discriminated easily.
- Since these phonemes are all used frequently by speakers who will participate in our experiments later, the pattern of the same phoneme can be generated naturally and with few differences at different time.

Besides, the duration of a whispered phoneme is also a potential discriminative feature which can tell whether it lasts long or short. (The duration threshold of short or long can be determined by analyzing user's individual habit or by a given value, e.g., 0.5 sec.) Therefore, each of the initial sivels corresponding to {/a/, /ə/, /i/, /ɒ/, /u/} can be pronounced in two forms, long and short, and considered as independent sivels. We use `a, e, i, o, u` to denote their short forms , and `A, E, I, O, U` for their long forms, and the sivel set selected empirically is $\mathcal{V}' =$ {`a, e, i, o, u, A, E, I, O, U`}.

C. Sivel Recognition

The speaker-dependent recognition experiments on samples of sivles in $\mathcal{V}'$ were performed to evaluate the effectiveness of the set of sivels as code elements.

The experimental data were collected from 6 speakers (1 female and 5 males). From each speaker, we recorded totally 500 samples (with 8KHz sampling rate) using our experiment system in office environment, that is, 50 samples of each sivel in set $\mathcal{V}'$. These 500 samples were divided into 50 groups, and each group contains one and only one sample for each sivel. Long-time spectral analysis were performed on the whole signal of each sample in order to get stable spectral feature (because these sivels are all monophthongs as we discussed above). Then each sample is represented by a 22-dimension parameter vector which contains 1 energy coefficient, 20 Mel-Frequency Cepstral Coefficients (MFCCs) and 1 duration coefficient. Linear Discriminant Analysis (LDA) were employed in training phase to reduce dimension of parameter vectors and select discriminative features. In testing phase, minimum Mahalanobis distance classifier was used as pattern classifier, labeling a testing sample with the class whose mean vector of training samples has the minimum Mahalanobis distance to that of the testing sample.

To evaluate the accuracy of speaker-dependent sivel recognition, 50-fold leave-one-out cross-validation (LOOCV) was performed on each single speaker's samples. For each run of the validation, one group of samples is used for testing while the rest groups of samples are for training. The final

 ISBN: 978-1-61208-195-3

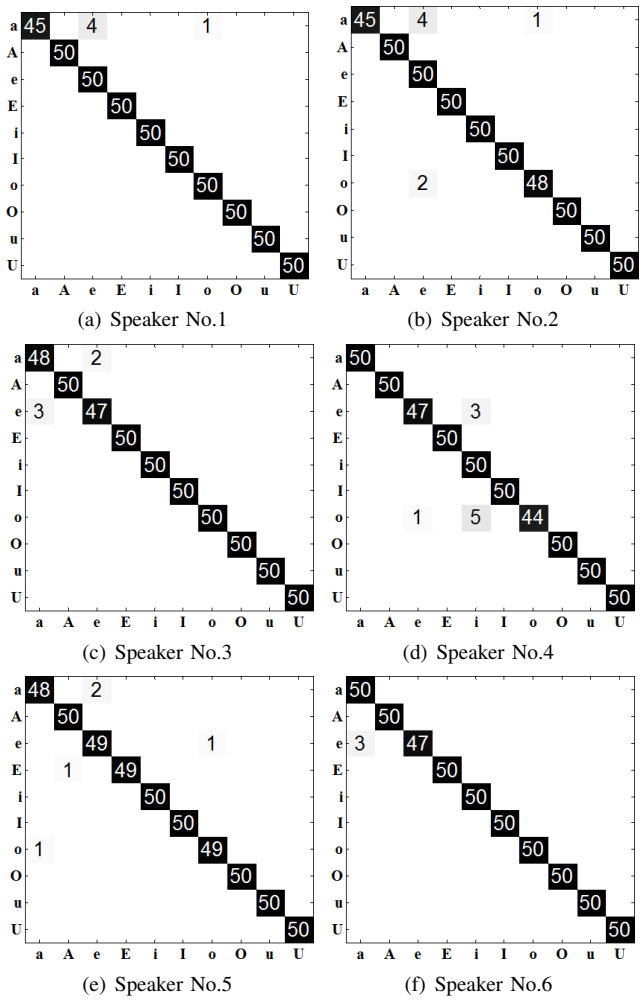


Figure 3. Confusion matrices of the speaker-dependent sivel recognition experiments on $\mathcal{V}'$ of 6 speakers, respectively.

accuracy is calculated by averaging accuracies of all the 50 runs. The experimental results for 6 speakers are illustrated in Fig. 3 and Table I. Observation of the results can be summarized and discussed as follows.

- The results of speaker-dependent recognition experiments all achieve high accuracies, and the mean accuracy is 98.87%. It means the sivels in set $\mathcal{V}'$ could be efficient code elements for text coding.
- There was no confusion between a short sivel and a long sivel, and among long sivels. This demonstrates that duration is a discriminative feature and it could not add confusions among the initial sivels.
- For each speaker, most errors of the recognition are caused by a few certain sivel pairs, such as the confusion between `a` and `e` in Fig. 3(a), 3(b), 3(c) and 3(f), and the confusion between `i` and `o` in Fig. 3(d).
- Every speaker may have their own pairs of confused phonemes in soft whisper due to their personal pronouncing habits (e.g., their accents).

Table I
ACCURACIES OF THE SPEAKER-DEPENDENT SIVEL RECOGNITION EXPERIMENTS ON V' FROM 6 SPEAKERS.

No.	Gender	Accuracy	No.	Gender	Accuracy
1	Female	99.0%	4	Male	98.2%
2	Male	98.6%	5	Male	99.0%
3	Male	99.0%	6	Male	99.4%
On average		98.87%			

We can see that sivels have the potential to be used as code elements for text input. However, the sivel selection method used in this section is not suitable for every speaker due to personal pronunciation habits. It is necessary to develop a general selection method to select a set of sivels for individuals.

III. SIVEL SELECTION

The sivel selection problem, selecting sivels from a number of phonemes, is described as follows: Given a N-cardinality set $\mathcal{P} = \{p_l | l = 1, \ldots, N\}$ of candidates, find a subset $\mathcal{V}$ which has high recognizability and a proper cardinality . The selection can be accomplished by calculating the recognizability of each subset of $\mathcal{P}$, then picking one subset that has acceptable recognizability and cardinality as the set $\mathcal{V}$, i.e., using so-called exhaustive method. The recognizability of a set is mainly dependent on its classification complexity which characterizes the difficulty of the classification problem on its elements' samples. A number of approaches have been used to measure the classification complexity, such as those mentioned by [10]. Since our aim is to recognize sivels, the LOOCV error rate is an appropriate measure for classification complexity , and further, for the recognizability as well. Unfortunately, the number of the subsets is huge due to the combination explosion, and LOOCV is time-consuming. If LOOCV error rate is used as the measure in the exhaustive method, the computational cost will not be accepted. To reduce the computational cost, we use a heuristic search method to find the set $\mathcal{V}$ of sivels.

As discussed above, most of the LOOCV errors on a set of phonemes are caused by a few certain pairs of its elements, and the pair with highest classification complexity is suggested to contribute the most negativity to the recognizability of the set. Therefore, define $E\big((p_i, p_j)\big)$ where $1 \leq i \neq j \leq N$ to calculate the one-on-one classification complexity of a pair of two different elements in $\mathcal{P}$, then for a subset $\mathcal{Q} \subseteq \mathcal{P}$ where $2 \leq |\mathcal{Q}| \leq N$, we use the maximum of $E(\cdot)$ on $\mathcal{Q}$ as a heuristic estimate of its holistic classification complexity. The heuristic estimate is denoted by

$$H(\mathcal{Q}) = \max_{q_i, q_j \in \mathcal{Q}} E\big((q_i, q_j)\big). \quad (1)$$

The values of $H(\mathcal{Q})$ for all possible $\mathcal{Q}$ while $|\mathcal{Q}| = 2, \cdots, N$, can be calculated in a recurrence way, i.e.,

$$H(\mathcal{Q}) = \begin{cases} E\big((q_1, q_2)\big), & \text{if } |\mathcal{Q}| = 2; \\ \max\Big(H(\mathcal{Q} - \{q\}), \\ \quad \max\limits_{q^- \in (\mathcal{Q} - \{q\})} E\big((q, q^-)\big)\Big), & \text{else} \end{cases} \quad (2)$$

where $q_i, q_j \in \mathcal{Q}$; q is an arbitrary element in $\mathcal{Q}$. We make an assumption that there are no equal values of classification complexity for different pairs of candidates in $\mathcal{P}$. Given a instance of $E(\cdot)$ and k as the desired number of sivels, we can select a set $\mathcal{V}$ of sivels from the set $\mathcal{P}$ of candidates using the sivel selection algorithm summarized in Algorithm 1.

Algorithm 1 Sivel Selection Algorithm

Input:

The set of candidates for sivels $\mathcal{P} = \{p_l | l = 1, \ldots, N\}$; k denoting the cardinality of $\mathcal{V}$, $2 \leq k \leq N$;

The function $E(\cdot)$ to evaluate the classification complexity of two candidates in $\mathcal{P}$, i.e., $E\big((p_i, p_j)\big)$ where $1 \leq i \neq j \leq N$.

Output:

The k-cardinality set $\mathcal{V}$ of sivels.

1: Calculate all the values of $E\big((p_i, p_j)\big)$, then calculate all the values of $H(\mathcal{Q})$ where $\mathcal{Q} \subseteq \mathcal{P}$, $2 \leq |\mathcal{Q}| \leq k$ using Eq. 2;
2: $\mathcal{W} \leftarrow \{w | w \subseteq \mathcal{P}, |w| = k\}$, $M \leftarrow 0$;
3: **repeat**
4: $\quad \mathcal{W}^* \leftarrow \{w^* | w^* = \arg\min H(w)\}$;
5: $\quad M \leftarrow M + 1$, $\mathcal{I}^M \leftarrow \bigcap\limits_{w^* \in \mathcal{W}^*}^{w \in \mathcal{W}} w^*$;
6: $\quad \mathcal{W} \leftarrow \{w | w \leftarrow w^* - \mathcal{I}^M, w^* \in \mathcal{W}^*\}$;
7: **until** $|\mathcal{W}^*| = 1$
8: **return** $\mathcal{V} \leftarrow \bigcup\limits_{m=1}^{M} \mathcal{I}^m$.

IV. Sivel selection Experiment

The experiment was conducted to evaluate the performance of our sivel selection method. The experimental framework is as follows. Given a N-cardinality set $\mathcal{P}$ of candidates for sivels, we selected three sets of sivels for each k ($k = 3, \cdots, N$). One set is selected using the proposed heuristic search method with $E(\cdot)$ to measure the classification complexity of every two candidates. Another set is selected using the exhaustive method with LOOCV error rate to measure the classification complexity of a subset of $\mathcal{P}$. The third set is selected also using the exhaustive method but with $E^+(\cdot)$, the generalization of $E(\cdot)$ for multiple candidates, to measure the classification complexity of a subset. The three sets of sivels are compared on their LOOCV error rates. In the LOOCVs, a same classifier is employed for both the sivel selection and the error rate comparison.

A. Data Collection

We pick candidates for sivels from those phonemes pronounced with fixed articulatory positions. In our experiment, there were 18 phonemes picked into set $\mathcal{P}$ as candidates for sivels, including

- 12 vowels: /ɑ/, /ʌ/, /æ/, /ɔ/, /i/, /ɪ/, /ə/, /ɜ/, /e/, /ʊ/, /u/ and schwa;
- 1 liquid: /l/;
- 5 fricatives: /f/, /θ/, /s/, /ʃ/, /h/.

Their samples were all collected from one male speaker. 60 samples of each phoneme were recorded by our experiment system with sampling rate of 8KHz in office environment. All the samples were divided into 60 groups in the similar way to that in the previous recognition experiment for 60-fold LOOCV. To avoid the influence caused by samples' duration, spectral analysis was only performed on the center 128-millisecond segments of each sample. Each sample is represented by a 21-dimension vector containing 1 energy coefficient and 20 MFCCs.

B. Experimental Configuration

The functions $E(\cdot)$ and $E^+(\cdot)$ are constructed to be positively correlated with the classification complexity. Fisher's discriminant ratio (FDR) is a classic measure of classification complexity for data with two classes [10]. Its form for individual feature values is defined as

$$f = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}, \quad (3)$$

where $\mu_1, \mu_2, {\sigma_1}^2, {\sigma_2}^2$ are the means and covariances of the two classes, respectively. Multiple features, e.g., d features, can form a d-dimensional column vector. Its form for feature vectors is defined as

$$F\big((p_{l_1}, p_{l_2})\big) = \max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{|\boldsymbol{w}^t \boldsymbol{S}_{\mathrm{B}} \boldsymbol{w}|}{|\boldsymbol{w}^t \boldsymbol{S}_{\mathrm{W}} \boldsymbol{w}|}. \quad (4)$$

$\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$ are between-class scatter matrix and within-class scatter matrix given by

$$\boldsymbol{S}_{\mathrm{B}} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^t \quad (5)$$

and

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{i=1}^{2} \boldsymbol{S}_i = \sum_{i=1}^{2} \sum_{\boldsymbol{x} \in \mathcal{D}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t, \quad (6)$$

where $\mathcal{D}_i$ is the vector set of the ith class; $\boldsymbol{m}_i$ is the mean vector of $\mathcal{D}_i$; $\boldsymbol{w}$ is a d-dimensional vector. Multiple discriminant ratio (MDR) is used as the generalization of FDR for data with multiple classes. It is defined as

$$F^+\big((p_{l_1}, \cdots, p_{l_c})\big) = \max_{\boldsymbol{W}} J^+(\boldsymbol{W}) = \frac{|\boldsymbol{W}^t \boldsymbol{S}_{\mathrm{B}}^+ \boldsymbol{W}|}{|\boldsymbol{W}^t \boldsymbol{S}_{\mathrm{W}}^+ \boldsymbol{W}|} \quad (7)$$

 ISBN: 978-1-61208-195-3

where c is the number of classes. The multi-class generalization of $\boldsymbol{S}_{\mathrm{B}}$ and $\boldsymbol{S}_{\mathrm{W}}$ is given by

$$\boldsymbol{S}_{\mathrm{B}}^{+} = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^t \qquad (8)$$

and

$$\boldsymbol{S}_{\mathrm{W}}^{+} = \sum_{i=1}^{c} \boldsymbol{S}_i = \sum_{i=1}^{c} \sum_{\boldsymbol{x}\in\mathcal{D}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t \ , \qquad (9)$$

where n_i is the cardinality of $\mathcal{D}_i$; $\boldsymbol{m}$ is the global mean vector of all the c classes. $\boldsymbol{W}$ is a matrix whose size is $d \times rank(\boldsymbol{S}_{\mathrm{B}}^{+})$. Since the values of FDR and MDR are negatively correlated with classification complexity, we constructed $E(\cdot) = -F(\cdot)$ and $E^{+}(\cdot) = -F^{+}(\cdot)$, respectively. Support vector machines (SVMs) were employed as the classifier in LOOCVs, and LIBSVM [11] with Gaussian kernel was used to implement SVMs.

C. Comparison Experiment and Results Analysis

For each given k ($k = 3, \cdots, 18$), three sets of sivels were selected using our method with FDR, exhaustive method with LOOCV error rate and exhaustive method with MDR, respectively. We computed the LOOCV error rate of each of the three sets, as well as the mean error rate of all the k-cardinality subsets of $\mathcal{P}$. Four curves corresponding to the four error rates by different k are drawn in Fig. 4 and named after "Ours+FDR", "Exhau.+LOOCV", "Exhau.+MDR" and "Mean Error Rate". The sets selected by the exhaustive method with LOOCV error rate can be taken as the best solution for the sivel selection problem, and the mean error rate can be seen as the effectiveness of a random solution. From the experimental results we can see that: our method with FDR usually gives the approximate optimal or the optimal solution, whereas the exhaustive method with MDR could not provide such excellent solution; all the error rate curves present upward trend as k is increasing, which means that more sivels will degrade the accuracy of sivel recognition though they can encode characters more efficiently (i.e., with shorter average code length). A tradeoff for a better global performance should be made between number of sivels and accuracy of recognition. After all, it can be summarized that our sivel selection algorithm has selected sets of sivels with low computational cost and the approximate optimal solution. These sets of sivels have rather low LOOCV error rate. With a proper k, sivels are efficient as code elements.

V. Sivel-based Text Input

The sivel-based text input adopts a customized scheme of encoding characters with sivels similarly to Morse code. The "characters" here include 26 letters, space, comma, period, digits 0∼9 and some control commands such as Backspace and Enter. A user can input text to machines by speaking relevant sivels.

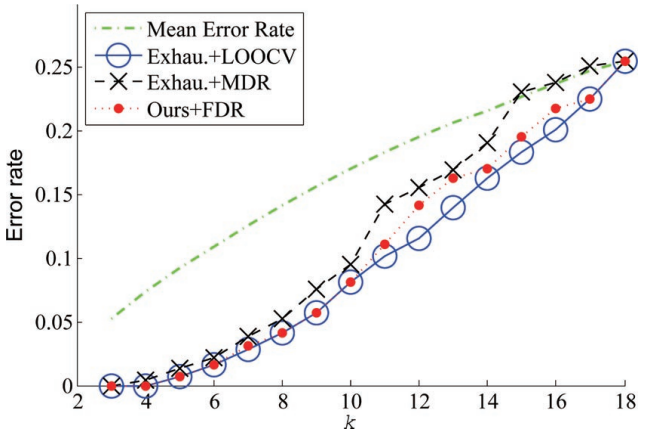


Figure 4. Comparison of the sets of sivels resulted by different methods on their leave-one-out cross-validation error rates .

We describe the usage of sivel-based text input by taking a user's input of a test sentence "hello world, i am a sivel typewriter." for an example. The user is the male speaker who participated in the sivel selection experiment. Using our proposed method, his personal sets of sivels as code elements are selected from the set $\mathcal{P}$ of candidates for sivels. 16 sets of sivels are resulted corresponding to different k ($k = 3, \cdots, 18$). Their 60-fold LOOCV (SVM as the classifier) error rates are computed and drawn as a curve by k, i.e., the curve denoted by "Ours+FDR" in Fig. 4. Owing to the "tradeoff" mentioned above, we choose the set selected when $k = 6$, $\mathcal{V} = \{$ /ɑ/, /ɔ/, /l/, /u/, /s/, /h/ $\}$ with the LOOCV error rate of 98.33%, as the set of sivels initially. Notice that the duration (how long a sivel's signal lasts) was not considered in the selecting phase, and as known from previous experiments, the duration can be used to discriminate between a short sivel and a long sivel. Therefore, the set $\mathcal{V}$ with 6 sivels is extended to the set $\mathcal{V}^{+}$ with 12 sivels by dividing each sivel into the short form and the long form. Similar to what we did previously, the extended set of sivels is denoted by $\mathcal{V}^{+} = \{$`a, o, l, u, s, h, A, O, L, U, S, H`$\}$. To evaluate the recognizability of these sivels, 50 samples for each of these 12 sivels were recorded from the user, and each sample is represented by a 22-dimension vector (1 duration coefficient added). 50-fold LOOCV on these samples achieved an accuracy of 99.67%.

Characters are encoded with these 12 sivels as Table II taking each character's appearance, usage frequency, pronunciation and each sivel's recognizability into account. Our rules are as follows.

- More frequently-used characters have shorter codes, such as space, backspace/enter, e, t, a, o, and i.
- Encode a character with sivel(s) pronounced as more similar to itself as possible, such as r, u, l, h, s, j, c, f, y, and v.

 ISBN: 978-1-61208-195-3

- Make codes look like the appearance of the character when they are put together, such as m, w, b, d, h, q, p, x, 8, k, and g.
- Use easier-to-pronounced combinations of sivels first, then that of easier-to-recognized ones, such as comma, period, z, n, and other digits.

Based on the text coding scheme above, the user can input text any character by speaking its sivel code. To recognize the stream of sivels spoken by the user, beforehand, the samples of these 12 sivels are used as training data of our experiment system. Then with the vibration sensor implemented on the position where the training samples were collected, the user is required to speak every sivel in isolation but with a short interval (within 0.5 sec) for each character. This is helpful to segmenting sivels and avoiding coarticulation effect. After speaking the code of a character the user pauses to wait for the feedback, and at the same time an interval longer than 0.5 sec is detected as the trigger of decoding. The resulted character is obtained and shown on screen (or its corresponding synthetic voice is sent to the user by earphones) as feedbacks. The whole sivel string spoken by the user is "h U l l o O uu o A l ol ou O S O a hh O a O s S uh U l O L ua la U uu A S L U A uo" in which a space means an interval longer than 0.5 sec. Ideally, it takes about 50 sec to input the test sentence with 37 characters by speaking 46 sivels. In practice, the cost time is around 63 sec with an input error rate of around 4%. The extra part of time is resulted by the user's reaction to feedback and the correction of input errors. To improve the inputting efficiency, an auto-correction strategy is used. This strategy allows the user to speak sivel codes of a word without waiting for the feedback of each character. After a space or a punctuation is input, the latest input word is automatically revised according to a dictionary. With the help of the auto-correction strategy, the user is able to input the test sentence within 55 sec.

VI. Discussion

Sivel-based text input (called sivel input for short) holds many advantages which is similar to speech input and NAM input (text input method with NAM recognition as its underlying recognition technology). They are all articulators-operated text input methods. They recognize the time series of code elements from signals generated by users' articulators, then generate text with a certain text coding scheme. Therefore, they can provide a useful channel for human-machine communication in some situations such as when users hands or eyes are busy, when hands-operated methods are difficult to be implemented, and when users can not move their arms or hands reliably due to disabilities. In addition, sivel input uses signals having little interaction with the ambient acoustic environment. This makes sivel input applicable to more situations with challenging acoustic environments than speech input such as where are noisy or silence-needed.

Table II
Characters and their codes with sivels in $\mathcal{V}^+$.

Char	Code	Char	Code	Char	Code	Char	Code
a	a	k	lu	u	u	4	UU
b	lo	l	l	v	uh	5	SS
c	ss	m	hh	w	uu	6	HH
d	ol	n	oh	x	uU	7	LL
e	U	o	o	y	ua	8	OO
f	hu	p	la	z	aH	9	OL
g	oa	q	al	0	AA	,	ou
h	h	r	A	1	aa	.	uo
i	S	s	s	2	oo	(Space)	O
j	sa	t	L	3	ll	(Bs/En)	H

The most significant difference among sivel input, speech input and NAM input is that sivel input adopts a customized scheme of text coding. Sivel input encodes text with only particular phonemes according to a customized scheme which can be optimized for specific users or tasks, whereas speech and NAM input encode text with conventional phonic units (phonemes and syllables) according to knowledge on linguistics. Although sivel input is not such a natural text input method as the other two, it is more effective in some special applications.

Here are two instances illustrating the advantages of the customized text coding scheme. One instance is the potential application of sivel input in secure communications. Encoding the messages with sivels itself is also an encryption process. Using sivel input, users can send messages quietly in various acoustic environments with high accuracy and without the participation of hands (and even eyes, if feedback via hearing), which makes the communication action difficult to be noticed by others. The other instance is that sivel input can enable speech disorder people to communicate with machines using an articulators-operated method. There are many speech disorder people who are not able to sound speech but only some phoneme-like segments of (silent) voice. They can select sivels from those segments of (silent) voice that they can sound reliably regardless of whether these voices have linguistic meanings. After giving user-customized names to these sivels and encoding text with them, speech disorder people are able to input text to machines using their articulators.

VII. Conclusion and Future Work

This paper has proposed the concept of silent voice elements (sivels) for text input. We selected a set of sivels empirically as an example, at first. The experiments of sivel recognition on the example show that sivels have potentials to be efficient code elements and they are speaker-dependent.

 ISBN: 978-1-61208-195-3

Then the selection of sivels for individuals has been accomplished using the sivel selection algorithm, and experimental results demonstrate that the algorithm can select set of sivels with high recognizability. We have introduced the sivel-based text input method in which characters are encoded with sivels according to a customized scheme. Using the sivel-based text input, a user inputted a sentence with 37 characters by speaking 46 sivels within 55 sec. Finally, the comparison among speech input, NAM input and sivel-based text input have been made and discussed to illustrate the advantages of the customized scheme adopted by the sivel-based text input.

For future work, we are currently testing the various robustness of sivel-based input, such as the robustness to the placement of the vibration sensors and how robust the sivel classifier are over time, and to improve sivel-based input for everyday utility.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and Prof. Petre Dini from IARIA for their valuable comments and suggestions to improve this paper.

This work was supported in part by the Natural Science Foundation of China(NSFC) under Grant No. 90920009 and NSFC-Guangdong Joint Fund under Grant No. U1035004.

REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 270–287, 2010.

[2] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical engineering & physics*, vol. 30, no. 4, pp. 419–425, 2008.

[3] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 341–353, 2010.

[4] M. Wester and T. Schultz, "Unspoken speech-speech recognition based on electroencephalography," Master's thesis, Karlsruhe: Universität Karlsruhe (TH), 2006.

[5] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *ICASSP*, vol. 5. IEEE, 2003, pp. V–708 – V–711.

[6] Y. Nakajima, "Development and evaluation of soft silicone NAM microphone," *IEICE Technical Report*, vol. 105, no. 97, pp. 7–12, 2005.

[7] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 301–313, 2010.

[8] V. Tran, G. Bailly, H. Loevenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *SPEECH COMMUN.*, vol. 52, no. 4, pp. 314–326, 2010.

[9] D. Babani, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic model training for non-audible murmur recognition using transformed normal speech data," in *ICASSP*. IEEE, 2011, pp. 5224–5227.

[10] T. Ho and M. Basu, "Complexity measures of supervised classification problems," *PAMI, IEEE Transactions on*, vol. 24, no. 3, pp. 289–300, 2002.

[11] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

 ISBN: 978-1-61208-195-3