Verification of phonological rules using automatic phone recognition

Tae-Yeoub Jang

Department of English
Hankuk University of Foreign Studies
270 Imun-dong, Dongdaemun-gu
Seoul 130-791, South Korea
tae@hufs.ac.kr

ABSTRACT

I introduce a way of verifying phonological processes on the basis of phonetic substances obtained by automatic speech recognition. The acoustic characteristics of phone-like units are modelled using automatic speech recognition techniques. This phone recogniser is run on the data tokens whose segmental structure matches with the context of target processes to be verified. Examining output strings of the recognition reveals whether the rule has been applied. Risk of erroneous recognition is alleviated as the final judgment is based on processing of multiple data tokens in a large spontaneous database.

1 INTRODUCTION

The proper representation of phonological processes are difficult. The traditional way of their formalism having been widely used in the framework of traditional generative phonology appears to be too strict since it frequently fails to describe variability of phonetic substances. This difficulty has also been a problem of Optimality Theory (OT, [1]), a more up-to-date trend of representing phonological processes in terms of constraint interaction, inevitably giving rise to an effort to a modified approach, within the same framework, based on statistical judgment of the processes ([2]). Indeed, some phonological processes are powerful enough to change the identity of a segmental unit whenever there is a matching context, while others are relatively flexible so that their application depends on other factors like speaker style or speed of the spoken utterances, producing various optional pronunciation variants.

I attempt to verify such processes using a method which I believe is more productive and objective. The basic idea can be summarised as: (a) when a target process to be verified is given, relevant speech tokens are collected which contain the environment of the rule. (b) A phone recogniser is run on each data token and produce an output string, and (c) the results are quantitatively analysed by investigating whether the target sound is affected or not. (d) Based on this analysis, the validity of the process can be determined. Each step will be described in detail through this paper.

An advantage of this approach is practicability. As the processes will be verified in terms of low level speech technology, they can be directly exploited for its improvement. For example, the processes can be used to generate plausible pronunciation variants of a language and subsequently enrich the pronunciation lexicon which is known for improving the performance of speech recognisers ([3], [4]). Besides, the current research, if found useful, is especially meaningful in the sense that speech technology is exploited in the linguistic phonetic validation.

2 DATA

The speech corpus used in this research is a British English spontaneous dialogue database known as the "Maptask" corpus [5]. It contains 128 dialogues, in each of which two participants converse based on the maps provided in advance. The database suits the aim of the current research as it contains voices of various styles and speech rates. The data tokens are divided into subsets to be used for different purposes. 60 dialogues are used for training phone recogniser, five dialogues for phone recogniser test, and another five dialogues for extracting tokens for rule verification. No subsets are overlapped. Each process will be described in detail in later sections.

3 PHONE RECOGNISER

To produce a surface phonetic form of each input speech, an automatic phone recogniser is developed

ISBN 1-876346-48-5

based on a well-known statistical technique, Hidden Markov Model (HMM). As for the recognition units I adopt a phone-like unit set designed by Centre for Speech Technology Research (CSTR) at the University of Edinburgh. The inventory consists of 57 phones and three silence units whose details are shown in [6]. Each phone model is represented by 39-dimensional parameter vectors consisting of features such as 12 Mel Frequency Cepstral Coefficients (MFCC), log energy, and their first and second order derivatives. These features are represented on left-to-right eight Gaussian mixture HMM's with three emitting states. The entire processing is conducted using the Hidden Markov Model Toolkit (HTK, version 3.0; [7]).

4 SELECTING TARGET RULES

4.1 RULE SELECTION METHODS

There are three ways to select target rules for verification. First, we can retrieve various rules by reviewing literature on phonology or phonetics. This is certainly the simplest way but a drawback is that only previously established rules are available. Nevertheless, it is important to count these rules in as the validity and practicality of them can be verified on purely phonetic basis. I referred to [8] and [9] to retrieve traditional phonological rules. Second, human investigation of phonetic substances can reveal many processes. A typical example is extraction of rules based on examination of phonetic annotations produced by speech labellers. When a phone string of a token is compared with the designated phonological form, the difference between two strings suggests the existence of a process regardless of whether it is systematic or frivolous. The advantage of this method is that lots of detailed phonetic phenomena, which have not been specified by previous phonological studies, can be discovered. The last method is also a phonetic identification of rules but this time the output strings to be examined are produced not by human investigation but by a speech recogniser. If a recognition system is constructed with phones as its basic units, it can be directly used to generate phone strings for each input speech. This means that a method based on comparison of output string with phonological string, mentioned just above, can be similarly employed. More detailed description on this bottom up approach of rule finding is shown in [6].

The rules found by these three ways are not of course mutually exclusive. Many rules selected on the basis of phonological theories (the first method) are also captured by the other two more practical methods.

I used all three methods in selecting target rules. However, as the current study is a preliminary effort to testify a new method of verifying rules, it deals with only a few post-lexical phonological rules of English.

4.2 TARGET RULES

Below are illustrations of target rules and word examples found by the selection processes described above.

Rules	Examples
Rotacism	[f a:] ⇔ [f a r] "far"
h-deletion	[h i m] ⇔ [i m] "him"
Glottalisation	[s i t i] ⇔ [s i ʔ i] "city"
Flapping	[s i t i] ⇔ [s i ɾ i] "city"
V Reduction	[æz] ⇔ [əz] "as"

The target rules shown in this section are only a small portion of phonological rules of the spoken English language. The base forms on the left side of each output pair are retrieved from a publicly available online pronunciation dictionary called BEEP (British English Example Pronunciation). Note that the base pronunciations specified are not abstract phonological representations by any means but phonetic entities which already underwent all the obligatory phonological rules. This suggests that the pronunciation comparison should be just a surface phonetic matter rather than phonological-phonetic mapping defined by UR-PR derivations of traditional theories of phonology. That is why I often use the term "process" in place of "rule". It implies that not only traditional phonological rules but also phonetic implementation rules and even constraints can be the target processes of the current study.

5 GENERATING PHONETIC OUTPUTS

Once target rules are determined, the verification of rules is straightforward. Running the phone recogniser on the selected data tokens will produce output phone strings for each token. A powerful pattern recognition technique known as *viterbi* algorithm [10], which computes the most likely state sequence, is used for the recognition process.

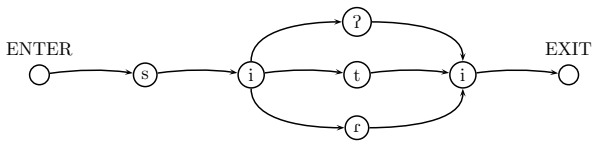
The most important work during the recognition process is setting up the collocational restrictions of phone sequences. A permissible phone network should be designed and implemented into every output token stream. If this procedure is omitted and a general phone sequence restriction is provided as is the case for usual phone (or other units) recognition process, frequent insertion of irrelevant phone items will be unavoidable and investigation of ouput string may need a separate procedure to eliminate those insertions and readjust the string either automatically or manually. Let's take the rule *rhotacism* applied to a word item "centimetres" as an example. When the phone recogniser was run on one of the tokens of this word, and when only the bigram language model was given without any further restriction, a resulted phone string was:

ISBN 1-876346-48-5

[silence e n ə n p i m i ə t ə u r ə silence]. Only with this output, it is almost impossible to match the string with the input word "centimetres". Nor the manual readjustment does not seem easy, if not impossible. Nonetheless, we can conjecture that there is a rhotic *r* at the relevant position: that is, word final syllable coda position.

Interested only in that specific position, and further more, in whether the rhotic *r* appears or not, we can stipulate the language model so that either of two strings [s e n t i m i: t ə r z], or [s e n t i m i: t ə z] is forced to appear phonetically. Likewise, another word "city" is forced to be realised as one of the three forms such as [s i t i], [s i ɾ i], and [s i ʔ i].

This manipulation is easily achieved by creating and implementing a simple finite state network which allows a N-ary transition path only at the relevant state. Below is an example of such artificially restricted finite state network search path for the word "city".



When the output is realised with the flap sound as in [s i ɾ i] the flapping rule is determined to have been effective. Likewise, the output as [s i ʔ i] means that the glottalisation rule has been applied.

6 VALIDITY JUDGMENT

The final step is running the recogniser on all the relevant data tokens for each of those target rules. As we are interested only in local acoustic information the relevant portions (usually words) of each speech data token is extracted and made to constitute an individual speech token. A uniform 50msec silence is inserted at the beginning and ending part of new tokens for the convenience of processing. Then the phone recogniser is run on these tokens as described in the previous section. Table 1 is the result of this processing for the target rules described earlier.

Results show that all the above phonological rules exist in the given data. The *flapping* rule has a relatively low application rate but it is still a considerable value. Two questions may arise regarding the judgment. First, what is the critical value to deny the existence of a process? Second, how can we trust the phone recogniser and safely say that the result values are not dominated by the recognition error?

	Ratio of Rule Application (%)	Number of Tokens
Rotacism	35	66
h-deletion	79	42
Glottalisation	25	43
Flapping	17	35
V Reduction	70	144

Table 1: Ratio of phone recognition outputs which underwent each process

To indirectly answer these questions a controlled recognition test has been conducted. 20 candidate words are arbitrarily selected for each rule. Half of them are tokens affected by each rule, while the others are tokens on which rule application is vacuous. This is to see how correctly speech recogniser automatically classifies the phonetic difference of phones in question. The results of this experiment can be summarised as:

	False Application	False Disregard	Accuracy (%)
Rotacism	2	3	87.5
h-deletion	3	0	92.5
Glottalisation	4	2	85.0
Flapping	2	3	87.5
V Reduction	7	4	77.5

The *false application* denotes outputs which are erroneously subject to the relevant rules. On the contrary, *false disregard* denotes outputs which are affected by rules when they are not supposed to. The worst accuracy is shown at the *vowel reduction* rule whose error rate is around 22.5%. That is, of 144 cases of vowel reduction in Table 1, maximum 34 cases may have been mis-classified. This statistic does not seem to be critical enough to attribute the total vowel reduction cases to accidents or errors. The outputs with respect to the other rules seem to be better identified in general by the speech recogniser. Consequently, we can carefully conclude, at this stage, the phone recogniser produces meaningful results though its performance varies depending on the type of individual rules.

7 DISCUSSION

Although the verification method proposed appears to be useful, there are restrictions. First of all, verifiable rules are required to be described in terms of recognition units, most frequently, phone-like units. In other words, when a speech recogniser does not differentiate allophones of a phoneme, there is no immediate means to test validity of the rule, as its output will always be a uniform phone symbol. For instance, the speech recog-

ISBN 1-876346-48-5

niser described above includes only one unit representing *voiceless bilabial stop*, such as [p]. Then, phonetic processes like aspiration and glottalisation, producing [p^{h}] and [p'] respectively, cannot be verified since they are not allowed to be phonetically produced in terms of separate units as recognised entity. The solution to this problem is straightforward. Detailed splitting of the unit seems enough. Of course, this solution entails other burdens of modelling acoustic characteristics of each individual phone, which could face a possible criticism. However, the increased complexity during the construction of recognition system is not crucial as far as the purpose of the work is just validation of rules, rather than optimisation of the ASR system itself.

Another problem is that the decision of the validity test is categorical since only designated allophones are possible outputs on which the validity judgment is based. Thus, when we attempt to see whether a word utterance underwent a specific rule, we can obtain the result of either 1 or 0, and nothing else. However, application of a rule need not necessarily be such categorical. Take a vowel devoicing rule as an example. While occurrence of a devoiced vowel is quite frequent in the context of surrounding voiceless sounds, many cases of relevant vowels are only partially devoiced. The proposed method of rule validation cannot capture this partiality phenomenon. Making partially devoiced phones as separate units and training them as individual phones is not an acceptable solution considering that degree of devoicing could be diverse. Nevertheless, the usefulness of the current method is not undermined critically. Having a number of data tokens in a database for each rule may alleviate this difficulty. Although the decision for each of the relevant data tokens cannot avoid being categorical, overall validation of a rule may not. For example, when all the categorical results for each individual voicing judgment are collected together, a final judgment on the rule can be made non-categorical as the ratio of the voicing will be calculated on the basis of scores for each class. If a non-optional rule does not have a severely biased ratio (say 60%-40% for voiced vs. voiceless), we can suspect that the rule applies only partially. Therefore, indirect judgment of non-categorical application of rules is not impossible.

8 CONCLUSION

A new method of verifying phonological processes is suggested. In spite of considerable recognition errors a phone recogniser is capable of producing legitimate phonetic strings with help of strict search path restrictions. Consequently, verification of phonological processes is available by examining output strings of the speech recogniser.

As many restrictions of the current approach are attributed to the restrictions of speech recognisers, enhancing recognition performance in general will result in better performance of rule verification. Further application of the proposed approach are also necessary by applying to other types of English phonological processes or even other languages.

ACKNOWLEDGMENTS

This paper is the modified version of [11].

REFERENCES

[1] Alan Prince and Paul Smolensky, "Optimality theory: Constraint interaction in generative grammar," Tech. Rep. 2, Rutgers University Center for Cognitive Science, 1993.

[2] Paul Boersma and Bruce Hayes, "Empirical tests of the Gradual Learning Algorithm," *Linguistic Inquiry*, vol. 32, pp. 45–86, 2001.

[3] H. Strik, J. M. Kessens, and M. Wester, Eds., *Modeling pronunciation variation for automatic speech recognition*, Rolduc, The Nethelands, 1998. European Speech Communication Association, University of Nijemegen.

[4] Tae-Yeoub Jang, "Fundamental frequency in manner differentiation of Korean stops and affricates," *Speech Sciences*, vol. 7, no. 1, pp. 217–232, March 2000, The Korean Association of Speech Sciences.

[5] Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth H. Boyle, Gwyneth M. Doherty, Simon C. Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert, "The HCRC map task corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[6] Tae-Yeoub Jang, "Identification and verification of english phonetic processes – a bottom up approach with a spontaneous dialogue database," in *Proceedings of the 2001 Acoustical Society of Korea Summer Conference*, July 2001, vol. 1, pp. 747–750.

[7] S. Young, J. Jansen, J. Ollason, and P. Woodland, *HTK Book*, Entropic, 1996.

[8] Philip Carr, *English Phonetics and Phonology – An Introduction*, Blackwell, 1999.

[9] Charles W. Kreidler, *The Pronunciation of English*, Blackwell, 1992.

[10] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, 1973.

[11] Tae-Yeoub Jang, "Phonetic verification of english phonological processes – a linguistic use of automatic speech recognition," in *International Conference on Speech Processing 01*, Seoul, Korea, 2001, vol. 2, pp. 981–985.