

Speech Recognition Based on Syllable and Pseudo-articulatory Features

Li Zhang and William Edmondson

School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK

L.Zhang@cs.bham.ac.uk

Abstract

The prevailing approach to speech recognition is the statistical technique known as hidden Markov modeling (HMM), which is capable of reasonable performance in general usage (~95%) – but not much more. The major drawback is that it ignores phonetics, which has the potential for going beyond the acoustic variations to provide a more abstract underlying representation. Also, HMM only produces a single interpretation of the speech data, i.e. a sequence of phonetic segments from acoustic waveforms. Other derived information such as syllables is based on the recovered segment sequence.

We have adopted a unique interdisciplinary approach to incorporate linguistic and articulatory knowledge into speech recognition processing. Additionally, our approach permits a “multithread” system because we recover syllable structure information from the speech stream directly without resorting to statistical models of segment sequences – and then use the syllable information to co-ordinate other interpretations of the data. The research will lead to a more plausible style of automatic speech recognition and will contribute to modeling and understanding of speech behavior.

1. Introduction

Currently HMM is the most popular approach to speech processing for recognition. It has achieved the best recognition results so far and with the help of powerful language models and careful dialogue design it is reliable enough to be implemented in commercial products. However, HMM relies heavily on the use of statistics to model the variability of speech, such as coarticulation effects and inter speaker differences, and the technique has nothing in common with the actual mechanisms of speech production or perception. Lee [1989], for example, states that “HMM is a very inaccurate model of the speech production

process”. It seems desirable to incorporate linguistic and articulatory knowledge into speech recognition systems, and recently speech recognition systems based on phonological and phonetic knowledge have gained more attention and interest. Such systems are more robust to noise, reverberation and speaker variability [Metze and Waibel 2002; Stuker et al. 2003].

Conventional HMM is mainly based on phones for modeling of spoken words. Any other recovered information such as syllables is based on this identification of phone sequences. However, not only have time and research shown that phones are too small an acoustical unit to model temporal patterns and variations in continuous speech, but also such processing is based on a single interpretation of speech data. This ignores information in the signal which can be used to derive syllable timing or structural information independently of phone sequences. More recently, attention has shifted to a larger acoustic context. Research has shown that speech recognition based on syllables can overcome some of the disadvantages caused by phone modeling systems [Hamaker et al. 1997; Ganapathiraju et al. 2001; Zhang 2003]. Finally, independent measures of syllable timing or structural information may provide more independent contributions and a speech recognition system based on this much richer base can make profound contributions to research on speech recognition processing.

Our novel approach to speech recognition is based on these three motivations. First of all, it incorporates more phonetic knowledge into the recognition process using pseudo-articulatory representations (PARs), which represent linguistic generalizations and idealizations of articulation and the articulator positions. This abstraction provides a powerful way to deal with problems such as articulatory-acoustic many-to-many mappings, and coarticulation [Iles 1995; Iskra 2000]. Secondly, our

approach shows that syllable structural information can be recovered directly from the speech stream in the form of PARs without any reliance on prior phonetic segment identification [Edmondson and Zhang 2001; Zhang and Edmondson 2002; Zhang and Edmondson 2003]. This allows speech recognition to proceed from a much richer base, which is important regardless of the detail of our approach. We believe that speech recognition systems must exploit in combination as many interpretations of the incoming signal as possible in a multithreaded style of processing. On the whole, this promises a more realistic approach to automatic speech recognition and will contribute to modeling and understanding of speech behavior.

2. Pseudo-articulatory representations

The approach we have taken is to develop a computational model for processing speech in a non-segmental way by using pseudo-articulatory representations. PARs can be described as the phonetician's idealizations of the articulatory process and are approximated by distinctive features in phonetics. Their values are, however, continuous rather than binary and range from 0 to 100. It has been demonstrated [Iles and Edmondson 1994] that in a simple case, and using PARs mapping formants to modified distinctive features taken from phonology, it is possible to overcome the ventriloquist effect, where acoustic evidence from many different articulatory configurations is recognized as a single phone. In general, PARs are abstract enough to discard the acoustic intricacies of the speech signal and the irrelevant fine details of articulation, and this makes them suitable for the work on recognition [Edmondson et al. 1996; 1999].

3. The syllable

The most prominent approach to speech recognition based on HMM is the use of phones for modeling of spoken words. However, time and research have shown that phones are too small an acoustical unit to model temporal patterns and variations in continuous speech. Thus, a need exists for a new technique capable of exploiting both the spectral and temporal characteristics of continuous speech. The focus has shifted to a larger acoustic context. The syllable is one such acoustic unit whose appeal lies in its close

connection to human speech perception and articulation, its integration of some co-articulation phenomena, and the potential for a relatively compact representation of conversational speech. Consequently, syllable-based modeling of speech has been used in speech recognition systems for languages that are considered more explicitly syllabic (e.g. Mandarin Chinese [Lee 1997] and Japanese [Nakagawa et al. 1999]), as well as in English-language speech recognition systems. Moreover, it is known that syllables are not only more stable in the speech signal compared to phonetic-segments [Ganapathiraju et al. [1997] but also play an important role in speech perception. Much prosodic information that is important for word recognition, such as stress-accent level and speaking rate, is directly tied to the syllabic representation of speech. All this suggests that syllable-level information should have a significant impact on speech recognition performance and it should be beneficial to model such syllable-related factors explicitly in automatic speech recognition systems.

3.1 *The articulatory pattern in the syllable*

There are several different ways of analyzing the syllable, and our first question is which is most useful as the basis for work on automated speech recognition? By comparing with several conventional ways, we have taken the approach which focuses instead on the notion that a syllable is basically an articulatory unit [Edmondson and Zhang 2002]. We have chosen to describe this, rather abstractly, as follows:

transition syllabic target transition

This expands to a more layered structure, shown in Figure 1, giving three layers altogether, where 's-tar' means syllabic target, 'd-tar' means dynamic target, 'tr-tar' means transition target, 'tr' means transition. The use of bold font in Figure 2 means that the identified component is marked for a specific 'phonetic' value, normal font means that the component is not identified as marked (it may have a complex specification, or no specification), italic means the component cannot be marked. Clearly, s-tar is always marked in reality (else there would be no syllable).

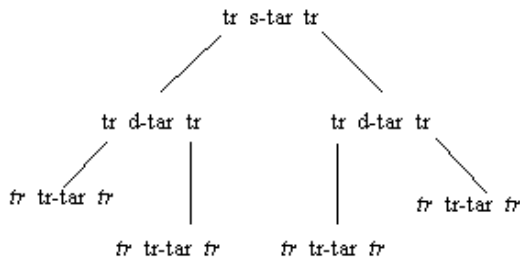


Figure 1 The layered syllable model

In this scheme articulatory activity must consist of *tr*, *x-tar*, *tr*, *x-tar*, *tr*, *x-tar* etc. where syllable nuclei are marked by *x* = *s*, and where phonetically irrelevant *tr* are *tr*. Typically, then, a CCCVCC syllable might look like:

tr, *tr-tar*, *tr*, *d-tar*, *tr*, *tr-tar*, *tr*, *s-tar*, *tr*, *tr-tar*, *tr*, *d-tar*, *tr*, *tr-tar*, *tr*

An example of how this might be used for the English word ‘apt’, is shown in Figure 2.

tr, **s-tar**, *tr*, **tr-tar**, *tr*, *d-tar*, *tr*, **tr-tar**, *tr*
 [æ] [>p] [pt] [t<]

Figure 2 An example of the syllable model

In figure 2, it shows that the articulatory detail can be labeled ‘phonetically’ but this does not equate to phones. The [p] is shown not as a phone, but rather just as the closure phase; likewise the [t] is shown as release phase. Additionally, complex articulatory activity, without phonetic significance but required for the phonetic string in which it is embedded, can be recorded, as in the case of the change in point of obstruction in the phase, or component, labeled ‘d-tar’.

4. Recognition

In the recognition process three successive stages can be clearly distinguished. The first stage is responsible for the transition from the acoustic representation of the incoming signal to the pseudo-articulatory representation with feature trajectories available as a function of time. The second stage concerns the movement from the pseudo-articulatory representations to the recovered syllable structures and produces a sequence of the recovered syllables. The third stage focuses on the transition from the syllable

patterns to the phonemic level of description and produces a sequence of phoneme labels. This third stage can be augmented by other phonemic labeling data derived using conventional techniques. This part of the work has been presented in detail in [Zhang and Edmondson 2002; Zhang 2003]. The final recognition results are very consistent and stable within every group of phonemes. The statistical analysis of the final recognition results is presented below.

5. Recognition results

Before moving on to the discussion of recognition results, the TIMIT database used in the processing, and also the results format are introduced and explained in order to help the reader to understand the result examples.

5.1 TIMIT database and results format

The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. The 10 sentences spoken by each speaker are divided into three groups: dialect, phonetically-diverse and phonetically-compact sentences. Our processing uses 10 sentences spoken by one speaker.

In order to explain the recognition results of recovered syllables here is one invented example.

| Record number (10msec intervals) | TIMIT phone symbol | Recovered syllabic detail |
|-------------------------------------|--------------------|---------------------------|
| 97 | A | s_tar, tr |
| 100 | B | tr_tar3 |

The waveform of every utterance is processed in 10msec samples, and the processing result of every 10msec is called a record. The record number is used to mark the records from the beginning to the end. In the TIMIT database, phone boundaries are presented by time. According to the relationship between the record number and the time, phone boundaries are presented by the corresponding record numbers instead. A or B represents the phone symbol as found in the TIMIT transcription file. In this example, phone B starts at record 98, and ends

at record 100. If the phone boundary in the TIMIT original transcription file does not coincide with the target boundary or the transition boundary in the recognized sequence, a number is placed after the target symbol or the transition symbol. This number refers to the number of records aligned with the particular original label. In the example, tr_tar3 in the recovered syllabic detail shows 3 records of transition targets are aligned with the original label B.

5.2 Recognition results of recovered syllable patterns

Our approach to speech recognition focuses on the use of syllable structure. The second stage of this approach concerns the derivation from the pseudo-articulatory representations of syllabic details and eventually produces a sequence of recovered syllables. Since this part is the core of our approach, we have presented below some examples of the results of recovered syllable patterns for the processing of 10 utterances. The transition targets (tr_tar) and syllable targets (s_tar) are very well recognized. The average accuracy rate for all the targets (including transition targets and syllable targets) is 72.8%, which is very promising.

We present one example from the processing in figure 3. On the left-hand side there are record numbers and original phone symbols as found in the TIMIT transcription files. Following the colon there are the recognized syllabic details. A crude time alignment has been attempted here. If the phone boundary, however, does not coincide with the target boundary or the transition boundary in the recognized sequence, a number is placed after the target symbol or the transition symbol. This number refers to the number of records aligned with the particular original label. This is why the numbers can be found only at boundaries. The correctly recognized targets (in accord with TIMIT) are printed in bold, i.e. where both the target and the time overlap.

5.3 Recognition results of final phoneme candidates

In order to evaluate the recognition results, we expand the phoneme labels over their duration. Therefore, if a phoneme is labeled to last 60 msec (whether it is the original utterance or the recognized one), it would be counted as 6 “occurrences” of the same phoneme (10 msec each). This is meant to evaluate not only the recognition of the phoneme, but to take into account its duration as well. Then a percentage is calculated by dividing the number of correctly recognized phonemes by the number of all occurrences of this phoneme in the original utterances. The recognition percentages are very consistent and stable within every group of phonemes. The vowels score highest, and among them the long vowel with 89% recognized correctly for /aa/. The nasals and semivowels follow with, e.g., 50% for /ng/. Stops are recognized pretty well, e.g. /bcl/ - 86%. Some of the fricatives are recognized pretty well too, e.g., /v/ - 67%, but other results are lower. On the whole, the fricatives and the affricates do not do very well.

```

Speaker ID: dr1/mmrp0
phonetically-diverse sentence
si2034: Make it come off all right.

si2034

15   h#:   s_tar, tr, tr_tar, tr
19   m:   tr_tar, tr
32   ey:  s_tar, tr, s_tar6
36   kcl:  s_tar4
40   k:   s_tar4
44   ix:  s_tar4
49   tcl:  s_tar5
50   t:   s_tar1
54   kcl:  tr, tr_tar, tr
59   k:   tr_tar, tr
65   ah:  tr_tar, tr
69   m:   tr_tar, tr
87   ao:  s_tar, tr2
97   f:   tr1, tr_tar, tr, s_tar2
109  ao:  s_tar11, tr
116  l:   tr_tar, tr, tr_tar,
124  r:   tr, tr_tar, tr, s_tar3
140  ay:  s_tar13, tr3,
148  tcl:  tr2, tr_tar, tr
156  h#:  s_tar

```

Figure 3 One result example of recovered syllable patterns

It is clear that some classes of sounds are recognized better than others, which is not unexpected. Vowels, semivowels and nasals have the best scores. These are the classes of sounds well known for their consistency, clarity and steadiness in their phonetic realization. These are also the sounds which can be described most adequately with the features selected earlier (high, back, etc.) [Iskra 2000]. Not surprisingly, the fricative sounds pose major problems, which is a case well known in automatic speech recognition and is due to the acoustic nature of these sounds. Therefore, future efforts to improve the recognition results will concentrate on this class of sounds.

The evaluation procedure used here is not optimal. The smallest chunk of labeled speech is 10 ms. Therefore, if the duration of a phoneme is, e.g., 57 ms, for the evaluation it would be assumed to stretch over 6 10-ms windows, the same as the phoneme with the duration of 63 ms. In reality, however, this difference could be quite significant and could account for some of the mistakes on the phoneme boundaries.

Finally, it is a well-established fact that in reality there are transitions between adjacent phones. As described in the syllable model in section 3.1, our approach identifies phonemes and transitions as final recognition results. Since the TIMIT database provides phones and phone durations one after another without any transitions, in order to make possible the comparison between the original transcription files in TIMIT, and our recognition results, we need to relabel the original transcription files in TIMIT to provide the phone sequence with reasonable transitions between them. However, one known source of errors in the existing system is the uncertainty surrounding timings and phone boundaries. The TIMIT database used in the work is known to have up to 25% incorrect indication of phonetic segment boundaries and this will contribute to recognition error scores. Moreover, the imposition of segment transitions into the TIMIT transcription files may be non-optimal, which may result in recognition errors as well.

6. Future work

On the whole, the phoneme recognition results can be regarded as promising. Introducing a few changes, such as improving phoneme models or correcting the time alignment problem, has the potential for ensuring more satisfactory recognition results which can then serve as the basis for introducing more speech data and more speakers. The approach offers a viable alternative for incorporating more phonetic knowledge into the recognition process, which is worthwhile pursuing further.

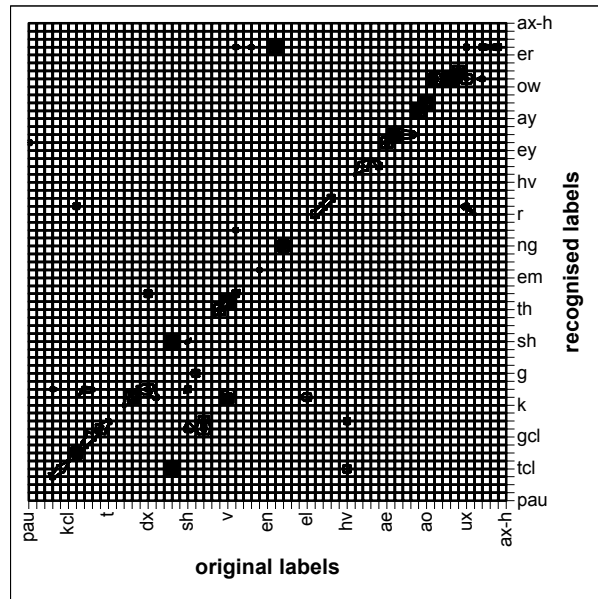


Figure 4 Some recognition results. The higher the recognition percentage, the darker the shading. Only some of the phoneme labels are visible. They are ordered in sound classes with silence/noise, plosives, affricates, fricatives, nasals, semivowels, and vowels from left to right and bottom to top.

7. Conclusion

Speech processing for recognition is conventionally concerned to recover a string of phones from the acoustic waveform. We have chosen here to explore the idea that it might be easier to recover strings of phonetically unlabeled syllables, and to use this information to recover phonetic detail without requiring that this detail be expressed in terms of phones.

We have also shown that it is possible to recover the desired details of syllables from speech without resorting to statistical models of phone sequences, or to models of the syllable as a sequence of phones. This suggests that the syllable is a good articulatory unit for speech recognition processing. Additionally, the work demonstrates the potential of processing speech to yield independent structures and characteristics, each of which can be assessed separately in terms of linguistic and articulatory plausibility, before being combined in a speech recognition system. We believe that speech recognition systems must exploit in combination as many interpretations of the incoming signal as possible. Other independent sources of information should also be added, for example timing and sonority, and future work will consider these possibilities. In general, we consider that this research will lead to a more plausible style of automatic speech recognition and will contribute to knowledge of phonetics and speech behavior.

8. Reference

- Edmondson, W., Iles, J. and Iskra, D. 1996. Pseudo-Articulatory Representations in Speech Synthesis and Recognition. International Conference on Spoken Language Processing (ICSLP 1996), 4:2215-2218.
- Edmondson, W., Iskra, D. and Kienzle, P. 1999. Pseudo-Articulatory Representations: Promise, Progress and Problems. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1999), 3: 1435-1438.
- Edmondson, W. and Zhang, L. 2001. Pseudo-Articulatory Representations and the Recognition of Syllable Patterns in Speech. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2001), 1:595-598.
- Edmondson, W. and Zhang, L. 2002. The Use of Syllable Structure for Speech Recognition. University of Birmingham. School of Computer Science. Technical Report CSRP-02-7.
- Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchoff, K., Ordowski, M. and Wheatley, B. 1997. Syllable – A Promising Recognition Unit for LVCSR. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, pp. 207-214, Santa Barbara, California, USA.
- Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G. and Picone, J. May, 2001. Syllable-Based Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing
- Hamaker, J., Ganapathiraju, A., Picone, J. 1997. Syllable-Based Speech Recognition. Technical Report. Prepared for Speech Research Group, Personal System Laboratory. Texas Instruments, Inc. Texas. Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University.
- Iles, J.P. and Edmondson, W.H. 1994. Quasi-Articulatory Formant Synthesis. Proceedings of ICSLP'94, 3:1663-1666.
- Iles, J. 1995. Text-to-Speech Conversion Using Feature-Based Formant Synthesis in a Non-Linear Framework. Ph.D. Thesis. School of Computer Science, University of Birmingham. UK.
- Iskra, D. 2000. Feature-Based Approach to Speech Recognition. Ph.D. Thesis. School of Computer Science, University of Birmingham. UK.
- Lee, K. 1989. Automatic Speech Recognition: the Development of the SPHINX system. Kluwer Academic Publishers, Boston.
- Lee, L. 1997. Voice Dictation of Mandarin Chinese. IEEE Signal Processing Magazine, (97): 63-101.
- Metze, F. and Waibel, A. 2002. A Flexible Stream Architecture for ASR Using Articulatory Features. Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002). Denver, USA. pp. 2133-2136.

- Nakagawa, S., Hanai, K., Yamamoto, K. and Minematsu, N. 1999. Comparison of Syllable-Based HMMs and Triphone-Based HMMs in Japanese Speech Recognition. Proceedings of the International Workshop on Automatic Speech Recognition and Understanding, pp. 197-200, Keystone, CO.
- Stuker, S., Metze, F., Schultz, T. and Waibel, A. 2003. Integrating Multilingual Articulatory Features into Speech Recognition. Proceedings of the 8th Eurospeech Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland. pp.1033-1036.
- Zhang, L. and Edmondson, W.H. 2002. Speech Recognition Using Syllable Patterns. Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002). 2: 1237-1240. Denver, USA.
- Zhang, L. and Edmondson, W.H. 2003. Speech Recognition Based on Syllable Recovery. Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland.
- Zhang, L. 2003. A Syllable-Based Approach to Speech Recognition Using Pseudo-Articulatory Features. Draft Ph.D. Thesis. School of Computer Science, University of Birmingham. UK.