Syllable Classification using Articulatory-Acoustic Features

Mirjam Wester

The Centre for Speech Technology Research
University of Edinburgh, Edinburgh
mwester@inf.ed.ac.uk

Abstract

This paper investigates the use of articulatory-acoustic features for the classification of syllables in TIMIT. The main motivation for this study is to circumvent the "beads-on-a-string" problem, i.e. the assumption that words can be described as a simple concatenation of phones. Posterior probabilities for articulatory-acoustic features are obtained from artificial neural nets and are used to classify speech within the scope of syllables instead of phones. This gives the opportunity to account for asynchronous feature changes, exploiting the strengths of the articulatory-acoustic features, instead of losing the potential by reverting to phones.

1. Introduction

In current state-of-the-art automatic speech recognition (ASR) systems, the acoustic signal is usually described in terms of phones, and words are simply seen as concatenations of phone sequences. However, as many before me have pointed out, the notion that a word is composed of a sequence of phone segments, i.e., the "beads-on-a-string" paradigm, is questionable [15, 7]. As articulators do not jump from one position to another, sounds also do not change abruptly, but the change is of a more gradual nature. This results in phenomena such as co-articulation and assimilation. Articulatory-acoustic features can be used to represent the acoustic signal in a compact way and they have a number of properties that make them an attractive solution to part of the "beads-on-a-string" problem, for instance, their asynchronicity and the fact that they can be used to represent co-articulation and assimilation effects as simple feature value changes.

The current study builds upon earlier work within the field of phonological features for ASR [1, 2, 3, 8, 9, 10, 11]. The phonological features employed in this paper are multi-valued features along different dimensions, i.e. place of articulation, manner of articulation and voicing as in [1, 11, 18]. The general approach to using articulatory-acoustic features in ASR is as follows. The starting point is acoustic training material with phonetic segment annotations. A mapping from the phones to the articulatory-acoustic features is carried out; i.e., each phone label is substituted by its corresponding articulatory-acoustic feature representation. Next, an array of artificial neural networks (ANNs) is trained on the basis of those articulatory-acoustic representations. One ANN is trained for each feature dimension. The input to the ANN is acoustic features, for instance, mel-frequency cepstral coefficients. The outputs from the ANN are estimated articulatory-acoustic feature values (posterior probabilities) for each 10 ms input frame. The output from the ANNs is used as input to train a new neural net [1, 11] or as the input features to train hidden Markov models (HMMs) [8, 9, 10]. This final step is carried out in order to perform the mapping from features to higher-level lexical units.

The studies mentioned above have all proved the viability of using articulatory-acoustic features as an alternative way of describing the speech signal. However, in none of the studies did using articulatory-acoustic features significantly outperform using conventional acoustic features.

One of the limitations of above named studies is that the step back to phones was always made, thus incorporating the "beads-on-a-string" paradigm back into the recognition process again. By going back to phones, the advantageous characteristic of the articulatory-acoustic features; asynchrony is not being employed. It has been shown that even when the neural networks are performing well, it is clear from the network output that articulatory-acoustic features often do not all change at phone boundaries [10]. Asynchronous feature value changes are common. Recognition models which are capable of modelling this asynchronicity properly should achieve significantly higher performance than the standard, frame synchronous systems. The observation that feature values do not change instantaneously at phone boundaries is echoed by findings in [2] which showed that not all frames are equal in terms of the ANN outputs. Some frames are classified more reliably, and it is those frames near the border between phones that are classified less reliably, which is further proof that there is no instant switch of feature values at phone boundaries.

Consequently, what is needed are methods that exploit the strengths of articulatory-acoustic features throughout the recognition process. To achieve this, we need to know how to combine the different feature streams while at the same time retaining the information that is present in the asynchrony of the different streams. I am convinced that an explicit link to syllable structure is essential in achieving this because:

- co-articulation is stronger within syllables than across syllable boundaries and co-articulation can be modeled by allowing phonetic features to overlap within syllables,
- articulations are generally programmed in syllabic units [5, 6],
- the asynchronous nature of the features can be better captured within a syllable than within phones.

Therefore, instead of going from articulatory-acoustic features to phones, I propose side-stepping the phones and going to syllables. The final goal of the present project is to improve the recognition results of ASR systems that are enriched with articulatory-acoustic feature information. This goal implies that methods to go from features to higher-level lexical units must be developed. This paper investigates how this can be achieved by investigating the syllable structure, and what position articulatory-acoustic features take in relation to syllables.

2. Material & Syllabification

This section describes the speech material that was investigated and explains how it was syllabified. Statistics on the syllables are also given.

2.1. Speech material

The speech material used in this study is from the TIMIT database [13]. TIMIT comprises hand labeled and segmented data of quasi-phonetically balanced sentences read by native speakers of American English. The drawbacks of using TIMIT for the present study are that TIMIT is read speech instead of spontaneous speech and, the test set contains a great deal of out-of-vocabulary words/syllables. Nevertheless, the fact that TIMIT has been manually transcribed weighs up to these drawbacks and as a starting point TIMIT is suitable for proving the validity of the proposed approach. All training sentences except SA sentences were used for training, and the core test set was used for evaluation. A random selection of 100 training sentences were used for cross-validation during training. The phone set used in TIMIT was reduced to 39 phones as in [14].

2.2. Syllabification

Syllabification software (tsylb2) available from NIST [4] was used to extract the syllables from the data. Two sets of transcriptions were syllabified, the manual transcriptions and canonical transcriptions. Canonical transcriptions of TIMIT were obtained by means of a dictionary look-up. A wraparound for tsylb2 was written to carry out the syllabification. The result is a time-aligned list of syllables in TIMIT format. A number of changes to the TIMIT transcriptions were necessary for tsylb2 to correctly parse the phone string. The main changes to the transcriptions were the merger of closures with the following burst, and the mapping of lone closures to stops. In addition, a number of final consonant clusters were added to the list of allowable final consonant clusters in tsylb2 because they occur in the TIMIT transcriptions. For example, /ng z/ was added to the final coda list as it occurs as the coda to "things" [th ih ng z] the standard being [th ih ng s].

2.3. Syllable statistics

In this section, various statistics about the syllables are given. The statistics were compiled on the basis of the output of the syllabification software.

Table 1 shows the number of unique words in the lexicon, the average number of variants per word, as well as the number of syllable tokens and types in the TIMIT training and test sets. Results are shown for both canonical and manual transcriptions of the training material. For the test set, only syllabification results for manual transcriptions are given.

Table 1: *Syllable statistics for TIMIT.*

data sets	**canonical training**	**manual training**	**test**
# words in lexicon	4891	4891	2372
average variants/word	1	2.2	1.9
# syllable tokens	47,301	47,397	17,227
# syllable types	3064	5525	3087

Further information that is of interest in this context is the number of syllables that are not present in the training material but nevertheless are present in the test material. The number of out-of-vocabulary (OOV) words and number of out-of-syllable-inventory (OOS) syllables are shown in Table 2. The final column in Table 2 shows the number of the OOS syllables that are not part of OOV words. This result indicates that a substantial portion of the OOS syllables are a result of OOV words. The remainder of the OOS syllables are a result of pronunciation variation, i.e. the transcription of the word in the test set does not match any of the examples of that word (or even parts of other words) in the training material.

Table 2: *OOV & OOS rates in TIMIT.*

	OOV words	**OOS syllables**	**OOS-OOV syllables**
number of OOV/S	1209	906	108
percentage of test set	26	9.6	1.1

Table 3 shows the consonantal-vocalic (CV) structure of syllables in the TIMIT training set. The percentage of training material covered by each syllable type is shown in the second column for the canonical transcriptions and in the final column for the manual transcriptions. There are three syllable types that occur in the manual transcriptions, but not in the canonical, and one that occurs in the canonical transcriptions but not in the manual. All of these, as Table 3 shows, are infrequently occurring syllables.

The first four lines in Table 3 are in accordance with previous findings [5]. "In spoken discourse, over 80% of the syllables are of the canonical CV, CVC, VC, V form, and many of the remainder reduce to this format by processes of assimilation and reduction [5]." In the TIMIT database, 78.3% of the canonical syllables have a CV, CVC, VC or V structure. When one considers the syllable structure of the manually transcribed data this percentage goes up to 82.4 %. This is of interest as a more simple syllable structure is easier to model.

Table 3: *Syllable types and coverage in TIMIT.*

syllable types	**canonical (%)**	**manual (%)**
[V]	7.10	11.00
[V C]	10.21	10.99
[C V]	30.84	36.04
[C V C]	30.17	24.41
[C C V]	4.54	4.87
[V C C]	1.83	0.83
[C C C V]	0.36	0.36
[C C C V C]	0.39	0.32
[C C C V C C]	0.09	0.06
[C C C V C C C]	0.02	0.00
[C C V C]	3.51	3.35
[C C V C C]	1.22	0.95
[C C V C C C]	0.12	0.07
[C V C C]	8.30	5.67
[C V C C C]	1.19	0.74
[C V C C C C]	0.02	—
[V C C C]	0.09	0.13
[V C C C C]	—	0.01
[C]	—	0.20
[C C]	—	0.02

3. Articulatory-acoustic features

In order to classify syllables, first the lower level building blocks, in this case, articulatory-acoustic features must be addressed. Table 4 shows the feature groups that were investigated and the values within each feature group. The target articulatory-acoustic feature representation was obtained by mapping from phonetic-segment labels to features. The mapping was based on [12] and very similar to the mapping pattern described in [2].

Table 4: *Articulatory-acoustic feature sets.*

feature	values
manner	approximant, fricative, nasal, stop, vowel, silence
voicing	+voice, -voice, silence
place	labial, labiodental, dental, alveolar, velar, glottal, high, mid, low, silence
rounding	+round, -round, silence
frontback	front, back, silence

Artificial neural networks (ANNs) were trained using the NICO Toolkit [17], which is an ANN toolkit designed and optimized for speech technology applications. For each of the five feature dimensions a separate ANN was trained. The architecture of the networks (similar to the architecture described in [16, 17]) was the same for all feature dimensions, with the exception of the number of hidden units and the number of output units. The number of hidden and output units for each feature dimension are shown in Table 5. The input units consisted of 12 Mel-frequency cepstral coefficients plus energy for 25 ms frames, with a 10 ms frame shift. In addition, deltas and double deltas were used. The connectivity from the hidden units to the output units was set to 25%, the connectivity from the input units to the hidden units was also 25% and the spread for the recurrent connections was 25. For more information on these parameters see [17]. The results in terms of percentage frames correctly classified are shown in Table 5. These results are in line with previously reported results [1, 9, 11]

Table 5: *Classification results for the articulatory-acoustic features.*

feature	% frames correct	# hidden units	# output units
manner	87.0	200	6
voicing	92.9	100	3
place	78.3	300	10
rounding	90.6	100	3
frontback	86.4	100	3

4. Feature Syllable Templates

Various studies [1, 6] have shown that there is a systematic relationship between articulatory-acoustic features and syllables. In [6], Greenberg explains how articulatory-acoustic features can give insight into the nature of pronunciation variation at the level of the syllable. A few of the points raised in [6] which are relevant to this study are the following;

- In a syllable, onsets are most often produced canonically, whereas the nucleus and coda are often reduced, the coda often even being deleted.
- Voicing is the articulatory foundation of the syllabic nucleus.
- It is rare for two segments of the same manner class to occur in adjacent positions within a syllable.
- Articulatory place cues serve to distinguish among words, particularly at onset. In coda position there is a general preference for central place of articulation.

These are all pointers to the type of cues that are present in speech material pertaining to the role of articulatory-acoustic features at the syllable level. The question that remains unanswered is how to extract this type of information and how to employ it in the speech recognition process.

As a starting point, to address this question, syllable templates were defined. The templates were derived from the manual transcriptions by rewriting the strings of phonetic segments in terms of articulatory-acoustic features and bundling them together at the syllabic level. The features for manner of articulation, place of articulation and voicing were considered. Note that describing the training and test material in this way, does not ensure unique descriptions for all syllables, in particular the nucleus is under-specified. Disambiguation of the nucleus (i.e. mainly vowels) will be done at a later stage, as prosodic prominence and lexical stress will have to play an important role to achieve this [1, 6]. An example of a number of templates is given in Table 6, which shows the syllable and template description for the words “ingenuity will”.

Table 6: *Example of syllable template descriptions.*

		template		
syllable	**type**	**voice**	**manner**	**place**
[ix n]	[V C]	+voi	vow_nas	high_alv
[jh ix]	[C V]	+voi	fric_vow	alv_high
[n uw]	[C V]	+voi	nas_vow	alv_high
[ax]	[V]	+voi	vow	mid
[t iy]	[C V]	-voi +voi	stp_vow	alv_high
[w l]	[C C]	+voi	appr	vel_alv

Figure 1 shows the same information as Table 6 in a graphical way. The syllabification shown in the first tier is derived from the manual phone transcriptions. The following three tiers show the target feature values for voicing, manner of articulation and place of articulation.

4.1. Classification of the syllable templates

In a classification task, the segment boundaries are known. In this case, the TIMIT syllable boundaries which were obtained during the syllabification process function as the segment boundaries. In addition to knowing the syllable boundaries, the syllable templates for all of the test material were also available, i.e. there were no OOS syllables. Activation values were obtained by running the networks for voicing, manner and place of articulation. These activation values were scaled and then normalized, ensuring the range for the outputs was between 0 and 1 and that the values summed to 1.

The probability for each syllable template for each feature group was calculated. A percentage correct was obtained by comparing the template with the highest probability to the correct template. For voicing, 78.1 % of the syllables were classified correctly, 64.8 % of the syllables were correctly classified in terms of manner of articulation and 53.2 % correct was found

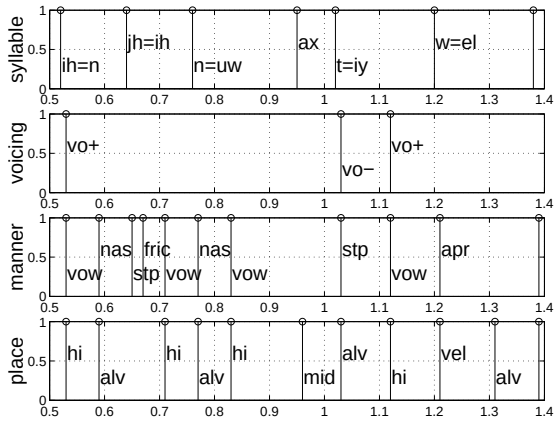

Figure 1: *Syllable template*

for place of articulation. Combining the templates for the three different feature groups to obain a syllable template leads to correct classification for 48.2 % of the syllables. These results show that more is needed than simple template matching to get from features to syllables, nevertheless, this gives a baseline to start from.

In ongoing work, weighted finite state transducers will be employed to learn the mapping from the articulatory-acoustic features to syllables. Other steps that will be taken are to include syllable position as an input parameter for the neural networks. In [1], significant gains are reported in both AF and phonetic classification accuracy when syllable-position information was incorporated in the neural networks.

5. Conclusions

The use of articulatory-acoustic features in ASR has been proved viable to a certain extent. Frame level accuracy rates are very high. However, the articulatory-acoustic features have not yet lived up to their full potential, as the frame accuracy rates do not translate into better recognition rates. This is due to the fact that their strong points have not yet been fully exploited in recognition experiments. It was argued that the asynchronous nature of the features is better modelled in syllables than at the level of phones.

In this study, an overview of the syllable structure in TIMIT was given. A first step in syllable classification was made by simple template matching. Future work will concentrate on employing WFST and alternatively decision trees to perform the mapping from feature streams to syllable templates. Syllabic prominence will also be incorporated in this framework to disambiguate between different syllabic nuclei.

6. Acknowledgements

The research described was supported by the Netherlands Organization for Scientific Research (NWO).

7. References

[1] S. Chang. *A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech Recognition*. PhD thesis, University of California, Berkeley, CA., 2002.

[2] S. Chang, S. Greenberg, and M. Wester. An elitist approach to articulatory-acoustic feature classification. In *Proc. of EUROSPEECH '01*, pages 1729–1733, Aalborg, 2001.

[3] S. Chang, L. Shastri, and S. Greenberg. Automatic phonetic transcription of spontaneous speech (American English). In *Proc. of ICSLP '00*, volume IV, pages 330–333, Beijing, 2000.

[4] B. Fisher. tsylb2-1.1 - syllabification software. http://www.nist.gov/speech/tools, August 1996.

[5] S. Greenberg. Understanding speech understanding: towards a unified theory of speech perception. In *Proc. of the ESCA Workshop on the "Auditory Basis of Speech Perception"*, pages 1–8, Keele University, 1996.

[6] S. Greenberg. Pronunciation variation is key to understanding spoken language. In *Proc. of ICPhS '03*, Barcelona, 2003.

[7] M. Huckvale. Exploiting speech knowledge in neural nets for recognition. *Speech Communication*, 9:1–14, 1990.

[8] S. King, T. Stephenson, S. Isard, P. Taylor, and A. Strachan. Speech recognition via phonetically featured syllables. In *Proc. of ICSLP '98*, pages 1013–1034, Sydney, 1998.

[9] S. King and P. Taylor. Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language*, 14(4):333–353, 2000.

[10] S. King, P. Taylor, J. Frankel, and K. Richmond. Speech recognition via phonetically-featured syllables. In *PHONUS 5: Proc. of the Workshop on Phonetics and Phonology in ASR*, pages 15–34, Saarbrücken, Institute of Phonetics, University of the Saarland, 2000.

[11] K. Kirchhoff. *Robust Speech Recognition Using Articulatory Information*. PhD thesis, University of Bielefeld, 1999.

[12] P. Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, 2nd edition edition, 1982.

[13] L. Lamel, R. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *DARPA Speech Recognition Workshop*, pages 100–109, 1986.

[14] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648, 1989.

[15] M. Ostendorf. Moving beyond the 'beads-on-a-string' model of speech. In *Proc. of IEEE ASRU Workshop*, pages 79–84, Keystone, CO., 1999.

[16] T. Stephenson. Speech recognition using phonetically featured syllables. Master's thesis, University of Edinburgh, 1998.

[17] N. Ström. Phoneme probability estimation with dynamic sparsely connected artificial neural networks. *The Free Speech Journal*, Issue #5, 1997.

[18] M. Wester, S. Greenberg, and S. Chang. A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proc. of EUROSPEECH '01*, pages 1729–1732, Aalborg, 2001.