interspeech 2008 incorporating SST 08

Fragmented Context-Dependent Syllable Acoustic Models

K. Thambiratnam, F. Seide

Microsoft Research Asia, 5F Beijing Sigma Center,
No. 49 Zhi Chun Rd., Beijing 100080, P.R. China
[kit,fseide]@microsoft.com

10.21437/Interspeech.2008-133

Abstract

Though touted as an excellent candidate, past work has yet to demonstrate the value of the syllable for acoustic modeling. One reason is that critical factors such as context-dependency and model clustering are typically neglected in syllable works. This paper presents fragmented syllable models, a means to realize context-dependency for the syllable while constraining the implied explosion in training data requirements. Fragmented syllables only expose their head/tail phones as context, and thus limit the context space for triphone expansion. Furthermore, decision-tree clustering can be used to share data between parts, or *fragments*, of syllables, to better exploit training data for data-sparse syllables. The best resulting system achieves a 1.8% absolute (5.4% relative) reduction in WER over a baseline triphone acoustic model on a Switchboard-1 conversational telephone speech task.

Index Terms: acoustic modeling, syllable, speech recognition

1. Introduction

Syllables are the basis of speech production and perception, a building block of language, and thus the correct unit for modeling speech. So cries the avid supporter of syllables, yet their defiant cries are silenced by vain attempts to outdo the well-established phone. Syllable approaches, to date, have made little gains over phone systems, and particularly when applied to real tasks such as conversational, unscripted speech. This is paradoxical, since the close ties between the syllable and articulation, and its ability to integrate co-articulation suggest that they should outperform the comparably naive triphone.

If all words were spoken as expected, there would be no reason to favour syllables over phones. Unfortunately, pronunciation in conversational speech is far from predictable, resulting in canonical pronunciations being a hopeless dream. There is however a level of predictability in syllables that makes them attractive. [1] found that the deletion rates for syllables was only 1% compared to 12% for phones. It was also found that the onset of syllables was generally well preserved in its canonical form, while the nucleus was almost always present though not necessarily in canonical form. A longer unit length provides for better discriminability and robustness for co-articulation. Since co-articulation effects can extend beyond the immediate neighbour phones, a syllable is able to better capture this co-articulation than a short-context triphone [2].

Fundamentally then, the syllable seems like an excellent modeling unit. In fact, syllable models are the model of choice in syllabic languages, such as Mandarin and Japanese. Syllable modeling in English though, is comparably more difficult due to a very large syllable vocabulary. Whereas Mandarin has a finite syllable set of 600 units (1500 with tones), English is said to have more than 10000 syllables. This results in data sparsity issues, since many of these syllables are rare. As a result, low-data syllables need to be backed-off to phones, leading to a mismatch in modeling paradigms. Additionally, in conversational speech, words are frequently resyllabified, leading to pronunciation variants and hence an increased number of units. Furthermore, even if only a small number of syllables are modeled, cross-word, context-dependent syllable modeling is intractable due to data sparsity. Clustering cannot be used to address this sparsity since the majority of syllables have barely enough data to train robust context-independent models. As a result, important methods used in triphone modeling, such as context-dependency and model clustering are ignored for syllable modeling.

This work addresses this issue be bringing context-dependency, cross-word modeling and model clustering to English syllable modeling, while constraining model explosion. This is achieved by fragmented syllable models, which enable context-dependent syllables while restricting surrounding models to using only the head/tail phones of a syllable as context. As a result, the number of context-dependent units is reduced compared to a naive context-dependent syllable system. Furthermore, fragmentation allows decision tree clustering to be applied, allowing sharing between syllables as well as phones.

The paper is organized as follows. Section 2 provides a brief introduction to syllable modeling and discusses key advantages and issues. Section 3 then introduces fragmented syllable models. This is followed by experiments and results in section 4 and finally conclusions and future work in section 5.

2. Syllable Modeling

A syllable can be defined as "a unit of spoken language consisting of a single uninterrupted sound formed by a vowel, diphthong, or syllabic consonant alone, or by any of these sounds preceded, followed, or surrounded by one or more consonants" [3]. They are often represented by their consonant-vowel structure, such as CVC, VC, CV and CCV. The central vocalic is called the nucleus, while the consonants before and after are referred to as the onset and coda respectively. This paper uses $x\hat{}y\hat{}z$ to represent a syllable, where x, y, z are the individual phones that make up a syllable eg. $b\hat{}ih\hat{}t$ is the syllable represented by the phone sequence $\{b, ih, t\}$.

Studies by [1] showed consistent behaviour with respect to the onset, nucleus, and coda, that make syllables attractive for modeling. By definition a syllable always has a nucleus, and this was reported to be realized almost always in real speech. However, nuclei can undergo deviations from their canonical forms, particularly for unscripted, conversational speech. The onset of a syllable was found to be generally well preserved in its canonical form when present in an utterance. In contrast, codas were found to be frequently deleted.

Accepted after peer review of full paper

September 22 – 26, Brisbane Australia

As a result, two types of variations can occur: here called inter-syllable and intra-syllable variation. Inter-syllable variation occurs for the same phone in different positions across syllables. For example, consider the realizations of **t** in "delicate idiot: $\{d, eh, l, ih, k, ax, \mathbf{t}, ih, d, iy, ih, t\}$" and "a **t**idbit: $\{ax, \mathbf{t}, ih, d, b, ih, t\}$". In the first, t is a unreleased syllable coda while in the second it is a stop-release syllable onset. A triphone model would fold both into the same triphone *ax-t+ih*. In contrast, they would be captured by different syllable models, $k\hat{}ah\hat{}t$ and $t\hat{}ih\hat{}d$ resulting in more robust models.

Intra-syllable variation occurs across realizations of the same syllable. Consider the sentence "i don't understand" - the t in "don't" is regularly dropped (called coda deletion). A syllable Hidden Markov Model (HMM) could capture this either implicitly by trained transition matrix probabilities, or explicitly by skip transitions [4]. Nuclei substitution is also common and could be captured implicitly in the state distributions (shown to be beneficial for pronunciation modeling by [5]) or explicitly through multi-pathing [6]. In contrast, triphones, could only indirectly capture this through a more complex state distribution.

Despite these benefits, prior works have shown only limited benefits and typically over sub-optimal baselines. [2] reported a 0.7% absolute gain over word-internal triphones, while [4] reported 0.5% absolute gain using a hybrid syllable-phone dictionary. A 0.5% absolute gain was reported by [6] for multi-path syllable HMMs. Finally [7] achieved a 0.5% gain over triphones using syllable feature tagging (stress, intra-syllable phone position and intra-word syllable position) of triphones.

2.1. Key issues: sparsity and context-dependency

Why are the gains from syllable modeling so moderate? One issue is data sparsity, a result of the large number of syllables (10000+) in English. The majority of these occur rarely, and thus it is not possible to train robust models. For example, for a 300 hour Switchboard-1 (SWBD-1) set, it was found that 79.6% of the 7946 syllables had less than 100 instances, 5.8% had more than 1000 instances and only 2.0% had more than 4000 instances. The low data syllables could be backed-off to phone models, giving a mixed syllable/triphone system, but this would reduce the syllable coverage of the data. For example, to ensure at least 4000 training instances per model, only 158 syllables could be used for SWBD-1 with a data coverage[1] of 68.9%, as shown in Figure 1. Furthermore, data allocated to syllables could not be used for phones unless models were trained independently resulting in less well-trained phones.

Another issue is context-dependent modeling, which requires more training data that is simply not available for the majority of syllables. Additionally, for cross-word contexts, the size of the recognition network is increased exponentially with the number of syllable units. Context-dependency is important for syllables, despite them being a more stable and longer unit than a phone. Effects such as coda deletion and nuclei substitution are cued by phone context, as are co-articulation effects.

3. Fragmented Syllable Models

Fragmented syllable models address the issues of data sparsity and context-dependency for syllable models. In a fragmented syllable model, a syllable is first represented by a state sequence such as a HMM. The syllable is then decomposed into three fragments: *left*, *centre* and *right*. This breakdown does not correspond to the typical onset-nuclei-coda breakdown. Rather the centre fragment, or c-frag, is chosen to cover states that are not affected by context ie. they are sufficiently well removed from co-articulation and other boundary effects. States before and after the c-frag states are allocated to the left (l-frag) and right (r-frag) fragments respectively. Of course, it is non-trivial to decide which states are non-context-dependent. Hence a simple heuristic is used, which is to allocate all but the first E and last E phones of the syllable to the c-frag. E here is termed the fragmentation extent. For syllables of length $\leq 2E$ phones, zero states are allocated to the c-frag.

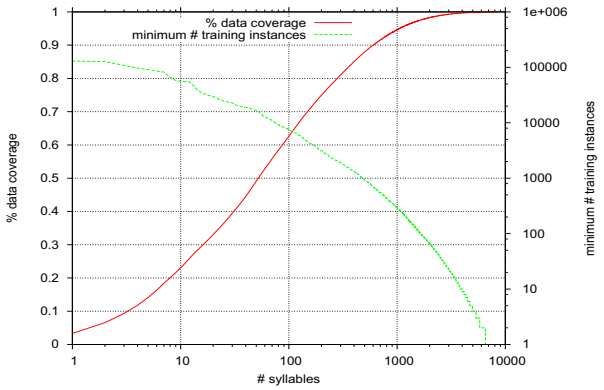

Figure 1: *Data coverage and min. # of training instances vs. number of syllable units*

C-frags are then labeled with the syllable identity, while the l-frags and r-frags are only labeled with the phones within their extent. For example, the syllable $b\hat{}aa\hat{}r\hat{}n$ and $E = 1$, would be fragmented as l-frag $b!$, r-frag $!t$ and c-frag $!b\hat{}aa\hat{}r\hat{}n!$. Note here the inclusion of a fragment marker, '!' which is used to differentiate between l-frags, r-frags, c-frags, and normal phones. Further examples are shown in Figure 2.

Triphone expansion can then be performed for all phones, l-frags and r-frags, but maintaining c-frags as context independent. Examples are shown in Figure 2. L/r-frags still preserve their syllable identity indirectly through context, and thus remain a part of the syllable model. However, syllables will not appear as context for phones, only l-frag and r-frags, thus constraining the context space. The reduction is inversely proportional to the fragmentation extent - for example, for $E = 1$ there was a reduction of 75% for the SWBD-1 300h training set, compared to a pure context-dependent syllable system.

In this way, a fully context-dependent model set is realized. Syllable states are distributed across multiple fragments, but still remain isolated from phone states. Both phones and syllables consider context: phones see only the first E phones of a syllable, while syllables see context via l/r-frags. Cross-word modeling is also possible since the growth of the number of units is equivalent to the number of unique l/r-frags, which is much less than the number of syllables.

3.1. Decision-tree clustering

Context-dependency however is data hungry, and thus these context-dependent models may be under-trained. Since c-frags are context-independent, data sparsity is less of an issue, as this can be controlled when choosing syllable models. However l/r-frags are context-dependent, and thus will share only a fraction of the total training data for a syllable. Sparsity is addressed by using decision tree clustering of l/r-frags that share the same centre phone, thus allowing data to be shared for low-data l/r-frags. Furthermore, l/r-frags can be shared across syllables, al-

[1] defined as number of syllable training instances after selection over number of training instances before selection

(a)	sil dh ey w^er m ae d sil	→ sil dh ey w! !er m ae d sil	→ sil sil-dh+ey dh-ey+w! ey-w!+!er w!-!er+m !er-m+ae m-ae+d ae-d+sil sil
	sil ey aa^n^d b iy sil	→ sil ey aa! !aa^n^d! b iy sil	→ sil sil-ey+aa! ey-aa!+!a^n^d! !a^n^d! !a^n^d!-!d+b !d-b+iy b-iy+sil sil
	sil ay d^ow^n^t n ow sil	→ sil ay !d !d^ow^n^t! !t n ow sil	→ sil sil-ay+d! ay-d!+!d^ow^n^t! !d^ow^n^t! !d^ow^n^t!-!t+n !t-n+ow n-ow+sil sil
(b)	sil dh ey w^er m ae d sil	→ sil dh ey w! !er m ae d sil	→ sil sil-dh+ey dh-ey+w! ey-w!+!er w!-!er+m !er-m+ae m-ae+d ae-d+sil sil
	sil ey aa^n^d b iy sil	→ sil ey a! !aa^n^d!n@2 !d b iy sil	→ sil sil-ey+aa! ey-aa!+!a^n^d!n@2 aa!-!a^n^d!n@2+d! !a^n^d!@2-!d+b !d-b+iy b-iy+sil sil
	sil ay d^ow^n^t n ow sil	→ sil ay d! !d^ow^n^t!ow@2 d^ow^n^t!n@3 !t n ow sil	→ sil sil-ay+d! ay-d!+!d^ow^n^t!ow@2 d!-!d^ow^n^t!ow@2+!d^ow^n^t!n@3 !d^ow^n^t!ow@2-!d^ow^n^t!n@3+!t !d^ow^n^t!n@3-!t+n !t-n+ow n-ow+sil sil

Figure 2: *Fragmentation examples: syllable → context-independent → context-dependent. (a) fragmented, (b) fully fragmented*

lowing even more data sharing, but this drops the constraint that l/r-frag states only belong to a single syllable.

3.2. Fully fragmented syllable models

Fragmented syllable models only allow model clustering for l/r-frags, not c-frags. Fully fragmented syllable models lift this restriction by allowing sharing for all fragment types. C-frags are further fragmented on a per-phone basis, but syllable identity is maintained on each fragment. These fragments are referred to as centre-phone fragments, or cp-frags. A syllable is now represented by a l-frag, followed by a sequence of cp-frags, and terminated by a r-frag, as show in Figure 2. Since cp-frags maintain syllable identity, they remain part of the syllable model and thus this fragmentation is not equivalent to the base phone sequence. Triphone expansion can then be done as shown in Figure 2, followed by decision tree clustering of cp-frags with the same base phone. For cp-frags with sufficient data, the decision tree will maintain separate states, while the remaining will be clustered. Thus data is shared across syllables at the cp-frag level allowing more robust low-data c-frags.

3.3. Clustered phone-syllable models

Fragments can be further exploited to share data between phone and syllable models. An unfortunate side-effect of syllable modeling is that training data is partitioned into two subsets: syllable-only and phone-only. Each group of models thus sees less training data. In fragmented syllables, l/r-frags can easily be clustered with phones with the same base-phone, as can cp-frags for fully fragmented syllables.

4. Experiments and Results

Experiments were performed on the SWBD-1 conversational telephone speech corpus. A training set of 300 hours with 39-dimension PLP cepstral mean/variance normalized features was used for acoustic model training. Syllabified transcripts were created using a state-of-the-art rule-based syllabifier.

The well-known 2-hour Eval2000 set was used for evaluation. A 22.6K vocabulary bigram language model trained on SWBD-1 and broadcast news data was used for decoding. It was not possible to use a trigram language model as certain experiments had model unit counts that exceeded the limitations of the available trigram decoder. Two baseline 40-mixture triphone HMM system were also maximum-likelihood trained. The first used a standard setup while the second used a position-dependent phone set, where the first phone in a word was tagged with *:s*, the last with *:e*, single-word phones with *:m* and all others with *:n*. To ensure a fair comparison, the total number of states (9300) and gaussians was kept the same across all baseline and syllable systems. The syllable configurations in Table 1 were trained. Syllables were bootstrapped by concatenating base phone sequence single-mixture monophone states. Fragmentation was then applied, followed by decision tree clustering and finally mixture incrementing to 40 mixtures. The number of parameters was consistent with the baseline models.

Name	Description
syl-ci-strict	Context-independent fragmented set. L/r-frags were syllable-internal biphones, instead of triphones. Equivalent to a standard context-independent syllable except phones use l/r-frags as context.
syl-ci	As *syl-ci-strict* system, except l/r-frags with different c-frag contexts were clustered ie. shared across syllables.
syl-cd-strict	Fragmented context-dependent, triphone l/r-frags, context-independent c-frags. L/r-frags with the same c-frag context (ie. same syllable) were clustered.
syl-cd	As *syl-cd* system, except l/r-frags with different c-frag contexts were clustered ie. sharing of l/r-frags across syllables.
syl-cd-share	As *syl-cd* but additionally sharing enabled between l/r-frags and phones.
syl-cd-full	Fully-fragmented context-dependent.
syl-cd-full-share	As *syl-cd-full* system but sharing enabled between l/r-frags, c-frags and phones.

Table 1: *Evaluated syllable configurations*

System	WER	
	Standard	Pos. Dep.
triphone	33.2	32.3
syl-ci-strict	55.4	55.0
syl-ci	38.1	37.7
syl-cd-strict	32.7	32.7
syl-cd	32.1	31.9
syl-cd-share	31.9	31.7
syl-cd-full	32.0	32.0
syl-cd-full-share	31.9	31.7

Table 2: *WER on Eval2000 set for syllable and baseline triphone systems for standard and position dependent phone sets*

Syllable unit selection was done by thresholding on the minimum number of training instances in the training data. Any syllables with less than M training instances were expanded to base phones. The majority of experiments used a minimum training instance count of 4000, which corresponded to 158 syllable models and a coverage of 68.9%. All fragmented systems used an extent of $E = 1$. For systems that required decision-tree clustering, a custom question set was created by adding l/r-frag and/or c/cp-frag questions. Fragment questions were adapted from both lexical group (eg. vowel, strident) and

System	Min. Train Instances	# Syls	% Train/Test Coverage	WER
syl-cd	800	546	88.0 / 86.8	32.5
	4000	158	68.9 / 66.5	31.9
	8000	88	57.5 / 54.7	31.8
syl-cd-full	800	546	88.0 / 86.8	32.2
	4000	158	68.9 / 66.5	32.0
	8000	88	57.5 / 54.7	31.8
syl-cd-full-share	800	546	88.0 / 86.8	31.4
	4000	158	68.9 / 66.5	31.7
	8000	88	57.5 / 54.7	31.7

Table 3: *WER on Eval2000 against syllable unit selection threshold for selected syllable configurations.*

single-unit context questions. Additionally, when clustering phones and fragments, questions on the triphone centre were added to allow separation of l/r/c-frags and phones.

Table 2 shows the Word Error Rate (WER) for standard and position-dependent phone sets. The context-independent *syl-ci-strict* syllable systems were the worst with WERs of 55.4/55.0% for standard and position-dependent respectively. This error rate was unexpectedly high, but consistent with similar results in [2]. Clustering l/r-frags resulted in an absolute gain of 17.3% for the *syl-ci*. This gain can be attributed to more data for l/r-frags as well as the ability to capture alternate co-articulations.

Adding context-dependency resulted in a WER of 32.7% for the *syl-cd-strict* systems, an absolute gain of 5.0% over the best context-independent system. Clustering l/r-frags gave a further 0.6/0.8% gain for the *syl-cd* systems. Context-dependency and model clustering clearly delivered notable benefits. Unfortunately, c-frag clustering did not appear to be beneficial. The fully fragmented *syl-cd-full* systems resulted in insignificant 0.1/-0.1% reductions in WER over the *syl-cd* systems. This is likely because, unlike l/r-frags, c-frags between syllables are acoustically unique and therefore do not benefit from clustering. This hypothesis is supported by the fact that syllables are stable and well-discriminable units.

There were, however, benefits from phone-syllable clustering. The *syl-cd-share* and *syl-cd-full-share* systems achieved gains of 0.2/0.2% and 0.1/0.3% respectively over their unshared counterparts, *syl-cd* and *syl-cd-full*. Phone-syllable sharing allowed the entire training data set to be observed by phone and syllable units, resulting in better trained models.

Overall, the best syllable systems were the *syl-cd-share* and *syl-cd-full-share* systems. Importantly, the position-dependent syllable systems achieved a 1.5% absolute gain (4.5% rel.) over a standard triphone system and 0.6% absolute gain (1.9% rel.) over a position-dependent triphone system. Additionally, they achieved absolute reductions of 23.5/23.3% respectively over standard context independent syllable setups.

Although accuracies for the *syl-cd-share* and *syl-cd-full-share* systems were equivalent, the latter is arguably a more convenient configuration since fully fragmented models have a homogeneous number of states. This homogeneity is typically exploited by optimized decoders and other algorithms. In fact, *syl-cd-full-share* is essentially a triphone system with appropriate tags attached to phones, and thus could be easily incorporated into any triphone configuration.

Further experiments were performed to examine the effect of amounts of training data per syllable. The results, as shown in Table 3, show that any WER disparity from increased syllable coverage is reduced for systems with more data sharing. For example, a 0.6% WER loss was observed for the 800-instance *syl-cd* configuration over the 4000-instance version. This system only used l/r-frag sharing. In contrast, only a 0.2% WER loss was observed for the 800-instance *syl-cd-full* system over the 4000-instance version, as a result of additional cp-frag sharing. For the *syl-cd-full-share* systems, a 0.3% WER gain was observed (a 1.8%/0.9% gain over the baseline triphone systems). Model clustering here allowed more data sparse syllables to be trained while mitigating (the quite small) WER impact.

4.1. Phone-syllable clustering

The decision tree of the 4000-instance *syl-cd-full-share* system was analysed to better under which models were being clustered. Clustered states that contained only a single type of unit: phone, l-frag, r-frag, or cp-frag, were labeled as pure, while others were labeled as mixed. Importantly, it was found that 48.2% of all states contained both phone and fragments. This was most likely a result of insufficient data to realize pure syllable states. It suggests that larger training sets could allow even better syllable models to be realized. Additionally, it was found that 41.5% of cp-frag states were isolated in pure states, while another 7.0% of were clustered with r-frags of shorter syllables. Furthermore, no two cp-frag states were found to be clustered together. These observations support the fact that syllables are acoustically unique units.

5. Conclusion

This paper has shown how fragmented syllable models can be used to realize context-dependent syllable acoustic models. The best fully-fragmented syllable system achieved a 1.8% absolute WER reduction over a standard triphone acoustic model and a 0.9% absolute reduction over a position-dependent triphone model. Additionally, a 23.3% absolute WER reduction was observed compared to a traditional context-independent syllable model. The reported experiments demonstrated the value of incorporating both context dependency and model clustering into syllable modeling. Nevertheless, the triphone system remains a hard baseline to outperform. Analysis of decision trees showed that many of the syllable units were still under-trained, and thus needed to share data with phone models. Future work will investigate whether more robust syllable models can be realized using even larger training sets.

6. References

[1] S. Greenberg, "Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation," in *Modeling Pronunciation Variation for Automatic Speech Recognition. Proceedings. ESCA Workshop on*, 1998.

[2] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *Speech and Audio Processing, IEEE Transactions on*, 2001.

[3] "The Free Dictionary," http://www.thefreedictionary.com/syllable, 2008.

[4] A. Sethy, B. Ramabhadran, and S. Narayanan, "Improvements in English ASR for the MALACH project using syllable-centric models," in *Automatic Speech Recognition and Understanding. Proceedings. IEEE Workshop on*, 2003, pp. 129–134.

[5] M. Saraclar, H. Nock, , and S. Khudanpur, "Pronunciation modeling by sharing gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, pp. 137–160, 2000.

[6] Y. Han, A. Hamalainen, and L. Boves, "Trajectory Clustering of Syllable-Length Acoustic Models for Continuous Speech Recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006.

[7] I. Shafran and M. Ostendorf, "Acoustic model clustering based on syllable structure," *Computer Speech and Language*, 2002.