INVESTIGATION OF ACOUSTIC UNITS FOR LVCSR SYSTEMS

X. Liu[1], *M.J.F. Gales*[1], *J.L. Hieronymus*[2] *& P.C. Woodland*[1]

[1]Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
[2] International Computer Science Institute
Berkeley, CA 94704, USA
Email: {*xl207,mjfg,jlh83,pcw*}*@eng.cam.ac.uk*

ABSTRACT

One important issue in designing state-of-the-art LVCSR systems is the choice of acoustic units. Context dependent (CD) phones remain the dominant form of acoustic units. They can capture the co-articulatory effect in speech via *explicit* modelling. However, for other more complicated phonological processes, they rely on the *implicit* modelling ability of the underlying statistical models. Alternatively, it is possible to construct acoustic models based on higher level linguistic units, for example, syllables, to explicitly capture these complex patterns. When sufficient training data is available, this approach may show an advantage over implicit acoustic modelling. In this paper a wide range of acoustic units are investigated to improve LVCSR system performance. Significant error rate gains up to 7.1% relative (0.8% abs.) were obtained on a state-of-the-art Mandarin Chinese broadcast audio recognition task using word and syllable position dependent triphone and quinphone models.

1. INTRODUCTION

One important issue in designing state-of-the-art LVCSR systems is the choice of acoustic units. Due to their well-defined and compact nature, context dependent (CD) phones have been the dominant form of acoustic units for well over a decade [4]. They provide a simple mapping between words and modelling units and good generalization to unseen words. Parameter tying techniques are used to further ensure robust estimation when only limited training data is available [2, 18]. CD phones capture co-articulatory effect in speech via *explicit* modelling. However, for many other more complicated phonological processes, they rely on the *implicit* modelling ability of the underlying acoustic models. This is often represented by the mixture models of tied HMM states. Alternatively, it is also possible to construct acoustic models based on higher level linguistic units, for example, syllables, to explicitly capture these complex patterns. Since large amounts of training data have become available, often in thousands of hours, this approach may show its advantage over implicit acoustic modelling, and thus has drawn increasing research interest in recent years [3, 7].

In this paper Mandarin Chinese is studied as an example to evaluate the performance of a wide range of acoustic units. Mandarin is a tonal language with syllabic structures represented by characters. It also shares several prominent phonological features with other languages. First, *tone sandhi* alters pronunciations of individual tonal syllables when they are adjoined to construct words [1]. A typical example is when there are two third tones occurring in immediate sequence, in which case the first tone changes to be very close to a rising, second tone. When there are more than three syllables in a word, the sandhi rules become more complex. Speaking style, dialect and accent can introduce further variability in the sandhi process. Second, *glottal stops* and the closely related *entering tone*, articulated at syllable end with a complete closure of the vocal tract [6], can also be found, particularly in accented speech. Finally, *stress* patterns act as alternations between syllables [5]. Sentential stress may diminish lexical level stress.

In order to capture the above phonological variability, three categories of acoustic units are investigated in this paper. The first uses longer context span phones to introduce stronger contextual constraints. The second employs word or syllable position dependent phones models; The third category uses explicit modelling of syllables. More complicated forms of acoustic units derived using a combination of them are also investigated. The rest of the paper is organized as follows. Phonetic decision tree based CD phone model clustering is reviewed in section 2. Position dependent phone models are presented in section 3. Syllable level acoustic models are proposed in section 4. Two issues when using these acoustic models are discussed in section 5. Experimental results on a state-of-the-art broadcast speech transcription task are presented in section 6. Section 7 gives the conclusion and suggests possible future work.

2. PHONETIC DECISION TREE CLUSTERING

Due to co-articulatory effects, the acoustic realization of the same phone can vary substantially depending on the surrounding phonetic contexts. To model such variation, CD phones are often used rather than monophones. One common form of CD phones is the *triphone* [2], which considers both the preceding and following phones. In order to incorporate more context information, it is possible to use CD phones with wider context span, for example, the *quinphone* or *pentaphone* [13]. A severe data sparsity issue arises when training CD phone models on limited amounts of training data. To handle this issue, parameter tying can be used to robustly estimate CD phone model parameters [18]. It may be flexibly performed at phone, state or Gaussian mixture component level, while state level tying is more commonly used in current LVCSR systems.

A range of clustering schemes have been proposed for state tying. The most widely adopted approach is to use a phonetic decision tree based clustering [18]. A phonetic decision tree is a binary tree with a set of ”yes” or ”no” questions at each node related to the context surrounding each base phone. Initially all states are clustered

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

at the root node. Splitting is performed by selecting the questions that locally maximize the likelihood of the training data. The change in log-likelihood when splitting a particular node p into D children nodes is approximated by [18]

$$\Delta \log p(\mathcal{O}) \;=\; -\frac{1}{2}\sum_{d=1}^{D}\gamma_d \log|\Sigma_d| + \frac{1}{2}\gamma_p \log|\Sigma_p| \qquad (1)$$

where $\mathcal{O}$ is the training data observation sequence, Σ_p and γ_p the covariance matrix and posterior occupancy of node p. Each tied state is also enforced to have a minimum amount of observed data. This ensures that rarely seen or unseen contexts are robustly handled. Finally, to generate a more compact tree, nodes are merged if the likelihood loss is below a threshold, until no such nodes are found.

Phonetic decision tree clustering provides a general and extensible framework for implicit modelling of phonological variability in speech. Richer linguistic constraints can be flexibly incorporated as additional questions to be considered during tree splitting [3, 7]. For example, for tonal languages like Mandarin, tonal questions can be asked during clustering to implicitly model different tonal variants of the same base phone [16].

3. POSITION DEPENDENT ACOUSTIC MODELS

As discussed in section 1, the occurrence of both tone sandhi and glottal stops bear a strong correlation with the precise position of particular phone or syllable within a whole word. One method to implicitly capture these patterns is to use position dependent (PD) phone models [4, 12, 15]. This involves splitting each tonal monophone into word or syllable initial, middle and final PD variants. The coupling of word and syllable positions can generate a maximum total of nine PD variants during decision tree clustering for each tonal base phone. As expected, some variants are invalid under the lexical constraint. Take the CU GALE Mandarin LVCSR system for example. Its baseline phone set consists of 46 toneless (124 tonal) phones derived by splitting some diphthongs in the original LDC phone set of 60 toneless phones. After incorporating word level position information, the the number of tonal phones is increased to 293. This is further enlarged to 487 by adding syllable level position information.

Right Front Stop	*+*_b, *+*_p
Right u2	*+*_u2
Left M_u2	*M_u2-*
Left fM_u2	fM_u2-*
... ...	... ...

Table 1. Section of position dependent question set.

Similar to the use of tonal information discussed in section 2, additional word level PD questions can be asked during decision tree clustering. This allows both word boundaries and phone positions to be explicitly modelled. It is also possible to further use each phone's position within a syllable, as an indirect way to model syllable structures. For example, **/fM_u2/** represents a tonal base phone **/u2/** that occurs in the middle and final positions of the whole word and syllable respectively. As only two post-vocalic consonants, **/n/** and **/ng/**, are normally allowed in Mandarin, syllable level position is more useful for vowels that may occur in any of three positions. An example section of PD question set is shown in table 1.

In order to increase the modelling power of PD phones, it is also possible to train PD models with wider context span. In this paper, both PD triphone and quinphone systems are investigated. A standard position independent (PI) triphone system, or a well trained PD system, may be used for initialization when clustering PD models. In practice, these two forms of initialization gave equivalent performance. In this work left-to-right HMMs with three emitting states and GMM densities are used for both PI and PD phone models.

4. SYLLABLE LEVEL ACOUSTIC MODELS

Syllables provide both a wider context span and stronger linguistic constraint than phones [10]. Explicit modelling of the three phonological features discussed in section 1 requires syllable structures to preserved. As discussed in section 3, the number of allowed post-vocalic consonants is small for Mandarin. Hence, it has a relatively compact set of base syllables. However, explicit modelling of all of them can still be problematic.

Base syllable cut-off (k)	**#Base units**	**#Tonal units**	**#CD units seen (M)**	**#CD units total (M)**
-	46	124	0.11	1.9
0	442	1487	1.32	3281
1	335	1226	1.31	1838
5	221	818	1.17	545
10	159	601	0.92	216
15	121	449	0.67	89.9
20	88	301	0.41	27.0
25	73	241	0.29	13.8
30	66	209	0.23	9.0

Table 2. Base syllable frequency cut-off, number of base/tonal units retained, CD hybrid units observed and the total of all possible units for a randomly selected 200 hour GALE Mandarin training set.

In the CU Mandarin system which uses a baseline 46 phone set (124 tonal), a cross word triphone context expansion gives 1.9 million distinct triphones, among which 0.11 million, approximately 6%, are observed in a 200 hour randomly selected GALE Mandarin training set. This is shown in the first line of table 2. In contrast, a phone to syllable level expansion of the lexicon gives a total of 442 toneless and 1487 tonal syllables, as is shown in the second line of table 2. A further tri-syllable context expansion leads to a colossal 3.3 billion distinct tri-syllables, among which only 0.4%, 1.32 million, are seen in the same data. For many rarely occurred base syllables, insufficient amounts of training data can result in poor modelling of surrounding contexts during decision tree based clustering. Hence, explicit modelling of all syllables is impossible.

To handle this issue, the most frequently occurred syllables found in the above 200 hour training set were merged with baseline phone set to give a combined set of hybrid phone-syllable units. As is shown in the second section of table 2, a range of base toneless syllable frequency cut-offs, from 30k to 1k were applied. Based on the performance of ML trained models after tying, the cut-off value at 25k was used, which retains a total of 27 toneless, and 117 tonal syllables. Combining them with the baseline phone set results in a 73 toneless and 241 tonal hybrid units, as is shown in the second line from bottom in table 2. During decision tree clustering, both phone and syllable context questions are asked to find the optimal splitting. For example, **u2-*** and **bˆu2-*** represent all CD phone-syllable hy-

brid units that have either tonal base phone **/u2/**, or syllable **/bˆu2/**, as the left context respectively. A more relaxed form, ***ˆu2-*** allows any syllable ended with tonal base phone **/u2/** to occur as the left context. An example section of the question set is shown in table 3.

Right Front Stop	*+b, *+bˆ*, *+p, *+pˆ*
Left u2	*ˆu2-*, u2-*
Left bˆu2	bˆu2-*
... ...	... ...

Table 3. Section of phonetic and syllabic hybrid question set.

Another issue in building syllable acoustic models is the appropriate model topology to use. In this work left-to-right HMMs with Gaussian mixture output distribution based state densities are used. The number of emitting states of each CD syllable is determined by three times the number of phones that the central syllable contains. For example, tri-syllable unit **bˆu2-rˆe4ˆn+wˆei2** will have a total of 9 emitting states. The advantage of this model topology is to allow initial alignment statistics used in decision tree clustering to be generated by a well trained tied triphone system, represented by triphone sequence **u2-r+e4 r-e4+n e4-n+w**. Finally, in order to fully capture the tone sandhi effect discussed in section 1, it is also possible to cluster and train word position dependent syllable models, in a similar fashion to the PD phones discussed in section 3.

5. ISSUES IN USING PD AND SYLLABLE MODELS

MPE error cost function: State-of-the-art LVCSR systems often use discriminative training techniques, for example, minimum phone error (MPE) training [14], as considered in this paper. An appropriate error cost function is required in MPE training. Using word and syllable position information, together with tones, significantly increase the number of phone classes when computing error costs. One important issue is thus the appropriate type of phone error to use for MPE. Three forms of error costs are evaluated in this paper, including using base toneless phones, tonal phones, or tonal PD phones. Using tonal phone labels only, but not position dependent information, was found to give the best character error rate (CER) for PD phone systems, as is shown in table 4. This was used in all experiments.

MPE Cost	bn06	bc05	d07	d08
phn	11.4	23.3	16.5	15.0
phn+tone	11.3	23.2	16.4	14.7
phn+tone+pos	11.4	23.4	16.5	14.8

Table 4. 3-gram LM unadapted MPE CER(%) using different error costs on a randomly selected 200 hour GALE Mandarin set.

Efficient use of contexts: As discussed in sections 3 and 4, the use of PD phones and syllables dramatically increases the number of CD acoustic units to consider during decoding. As not all of them are allowed by the lexicon, it is possible to consider using only the valid subset. One method to achieve this is to use a weighted finite state transducer (WFST) [11] *composition* operation, $C \circ L$, between the CD transducer C, which converts CD phone sequences to CI ones, and the lexicon transducer L, which maps CI phone sequences to words. The input symbol set of the resulting transducer contains all possible CD units allowed by the lexicon. This approach was used for all triphone and tri-syllable systems. It significantly reduced the number of CD units used at decoding time, for example, by 79% for the baseline triphone system, 98% for word and syllable PD triphone sysetms, and a less dramatic 60% for tri-syllable systems due to the increased connectivity of the syllable lexicon. These are shown in table 5. For quinphone systems a dynamic on-the-fly context expansion algorithm was used in decoding [13].

Context	Position	#Context total (M)	#Context active (M)
tphn	-	1.9	0.4
tphn	wd+sy	115	1.7
tsyl	-	13.8	5.5

Table 5. Number of context dependent phone or syllable units before and after intersection between CD and lexicon transducers.

6. EXPERIMENTS AND RESULTS

The CU Mandarin LVCSR system was used to evaluate the performance of various acoustic units. The full system was trained on 1960 hours of broadcast speech data. A total of 3.7 billion words from 28 text sources were used in LM training. A 63k word list was used. Five GALE Chinese speech test sets of mixed broadcast news (BN) and conversation (BC) genre: 2.6 hour **d07**, 1 hour **d08**, 3 hour **d09s**, 2.6 hour **p2ns** and 1.5 hour **p3ns** were used. The system uses a multi-pass recognition and system combination framework. It consists of an initial lattice generation stage using adapted baseline PI triphone MPE models and a multi-level LM followed by "P3" acoustic model re-adaptation and lattice rescoring stage before CNC combination. A more detailed system description can be found in [8].

Sys	Cntx	Position	d07	d08	d09s	p2ns	p3ns
a	tphn		12.4	10.6	12.6	11.8	16.1
b	qphn	-	12.1	10.6	12.6	11.6	15.9
c	tsyl		11.9	10.4	12.2	11.1	15.3
d	tphn	word	11.7	10.2	11.9	11.1	15.4
e	qphn	word	11.8	9.9	12.0	11.1	15.4
f	tsyl	word	11.8	10.1	12.2	11.3	15.2
g	tphn	wd+sy	11.7	10.2	11.8	11.3	15.4
a+b			11.6	10.2	12.2	11.3	15.3
a+c			11.7	10.1	11.0	11.0	15.0
a+b+c			11.4	9.9	11.7	10.7	14.9
d+e			11.4	9.6	11.7	10.6	14.8
d+f			11.4	9.9	11.7	10.8	14.8
e+g			11.4	9.6	11.6	10.7	14.9
d+e+g			11.1	9.7	11.4	10.5	14.7
d+e+f			11.0	9.4	11.3	10.6	14.4

Table 6. P3 and combination performance of baseline PI, PD phone and syllable based acoustic models trained on a 202 hour subset.

Initially a range of acoustic modelling units were evaluated on systems with 6k tied states, 16 Gaussians per state and MPE trained on a randomly selected 200 hours of data. Table 6 shows their CER performance. Using syllable units gave 0.2%-0.8% CER gains over the PI triphone system, but not the PD triphone baseline. For both

triphone and quinphone word level PD systems, consistent CER reductions of 0.3%-0.8% over the comparable PI baseline triphone or quinphone models were obtained. The CNC combination between triphone and quinphone models were also improved by 0.2%-0.7% absolute using word level PD models "d+e" over the baseline 2-way PI triphone and quinphone combination "a+b". The use of both word and syllable level position information (system "g") gave no improvement over using word level position only (system "d") for triphone systems, consistent with the results in [9], though it was found be useful in system combination with more diversity. The best combination performance were obtained using a 3-way combination "d+e+f" between word level PD triphone and quinphone systems, and a word level PD tri-syllable system, as is shown in the bottom line of table 6. Using this system, the overall CER gains over the baseline 2-way PI triphone and quinphone combination "a+b" are 0.6% (5.2% rel.) on **d07**, 0.8% (7.8% rel.) on **d08**, 0.9% (5.9%-7.4% rel.) on **d09s** and **p3ns**, and 0.7% (6.2% rel.) on **p2ns**. The genre specific absolute CER gains are 0.2%-0.6% for BN and 0.9%-1.4% for BC respectively.

Sys	Cntx	Position	d07	d08	d09s	p2ns	p3ns
a*	tphn		8.8	8.2	9.5	8.7	11.9
a	tphn	-	8.8	8.0	9.4	8.6	11.9
b	qphn		8.8	7.9	9.3	8.6	11.7
d*	tphn		8.6	7.9	9.0	8.3	11.3
d	tphn	word	8.5	7.8	8.9	8.2	11.3
e	qphn		8.4	7.8	9.0	8.1	11.2
a+b			8.5	7.7	8.9	8.3	11.4
d+b			8.5	7.6	8.8	8.2	11.2
a+e			8.4	7.6	8.8	8.2	11.2
d+e			8.4	7.7	8.7	8.0	11.1

Table 7. P3 and combination performance of baseline PI and PD phone acoustic models on the 1960 hour full training set.

Several word level PD triphone and quinphone systems were then trained on the 1960 hour full training set with 36 Gaussians per state. These are shown in table 7. First, a PD triphone system "d*" with 9k tied states was evaluated against its comparable 9k state PI baseline "a*". Compared with the 200 hour subset results in table 6, slightly reduced CER gains of 0.2%-0.6% absolute were obtained across all five test sets. The performance gains on BC genre were 0.2%-0.8%, and on BN genre 0.1%-0.3% absolute. The same performance improvements were also observed on a larger 12k tied state PD triphone system "d" against its comparable 12k PI baseline "a". These trends suggest position information can not be implicitly learned by standard PI phone models through having more clustered states. Similar CER gains of 0.1%-0.5% absolute were also obtained using a 12k tied state PD quinphone system "e" against its comparable PI quinphone baseline "b". Consistent but smaller gains up to 0.3% absolute were obtained in 2-way PD triphone and quinphone CNC combination "d+e" over the baseline PI triphone and quinphone combination configuration "a+b".

7. CONCLUSION

In this paper position dependent and syllable based acoustic units were investigated to model several prominent phonological features of Mandarin Chinese, and to improve LVCSR system performance. Experimental results on a state-of-the-art speech recognition task suggest word level position information may be useful to capture additional phonological variability in speech. Future research will focus on using additional prosodic information, such as stress condition, to improve syllable based acoustic modelling.

8. REFERENCES

[1] M. Y. Chen (2000). Tone Sandhi Patterns across Chinese Dialects, *Cambridge Studies in Linguistics*, No. 92.

[2] Y-L Chow & R. Schwartz et al. (1986). The Role of Word-Dependent Coarticulatory Effects in Phoneme-Based Speech Recognition System, in *IEEE ICASSP1986*.

[3] C. Fuegen & I. Rogina (2000). Integrating Dynamic Speech Modalities into Context Decision Trees, in *Proc. Eurospeech'00*.

[4] J-L. Gauvain, L. Lamel & G. Adda (2002). The LIMSI Broadcast News Transcription System, *Speech Communication*, Volume 37, pp. 89-108, 2002.

[5] G. Kochanski et al. (2003). Quantitative Measurement of Prosodic Strength in Mandarin, *Speech Communication*, Volumen 41(4), November 2003.

[6] P. Ladefoged (2005). Vowels and Consonants 2nd edition, Blackwell.

[7] H. Liao et al. (2010). Decision Tree Clustering with Word and Syllable Features, in *Proc. Interspeech'10*.

[8] X. Liu, M. J. F. Gales & P. C. Woodland (2010). Language Model Cross Adaptation For LVCSR System Combination, in *Proc. Interspeech'10*.

[9] J. Luo, L. Lamel & J-L Gauvain (2009). Modeling Characters versus Words for Mandarin Speech Recognition, in *Proc. IEEE ICASSP2009*.

[10] M. Jones & P. C. Woodland (1994). Modelling Syllable Characteristics to Improve Large Vocabulary Continuous Speech Recognition, in *Proc. IEEE ICASSP1994*.

[11] M. Mohri & M. Riley. Network Optimizations for Large Vocabulary Speech Recognition. *Speech Communication*, 25:3, 1998.

[12] D. B. Paul (1991). The Lincoln Tied-Mixture HMM Continuous Speech Recognizer, in *IEEE ICASSP1991*.

[13] J. J. Odell (1995). *The Use of Context in Large Vocabulary Speech Recognition*, PhD thesis, Cambridge University Engineering Department, England.

[14] D. Povey & P. C. Woodland (2002). Minimum Phone Error and I-smoothing for Improved Discriminative Training, in *Proc. IEEE ICASSP2002*.

[15] I. Shafran & M. Ostendorf (2003). Acoustic Model Clustering Based on Syllable Structure, *Computer Speech and Language*, 2003.

[16] R. Sinha et al. (2006). The CU-HTK Mandarin broadcast news transcription system, in *Proc. IEEE ICASSP2006*.

[17] P. C. Woodland et al. (1995). The 1994 HTK Large Vocabulary Speech Recognition System, in *Proc. IEEE ICASSP1995*.

[18] S. J. Young, J. J. Odell & P. C. Woodland (1994). Tree-based State Tying for High Accuracy Acoustic Modeling, *ARPA Human Language Age Technology Workshop*, pp. 307–312, Morgan Kaufman, 1994.