TWO-STREAM MODELING OF MANDARIN TONES

Frank Seide and Nick J.C. Wang

Philips Research East-Asia, Taipei

24FB, 66, Sec. 1, Chung Hsiao W. Rd., Taipei, Taiwan

{Frank.Seide,Nick.Wang}@philips.com

ABSTRACT

Tone modeling is a critical component for Mandarin large-vocabulary continuous-speech recognition systems. In previous work on pitch-feature extraction, we reported character error rate reductions of over 30% over the non-tonal baseline [1]. In this paper, we investigate how best to integrate tone modeling with a Mandarin LVCSR system.

The paper focusses on the *two-stream* method, which is based on two-stream continuous-mixture HMMs. For tonal languages, sub-word units may depend on both phonetic context and tone. To alleviate for the multiplication of model parameters, the two-stream method models state emission distributions as products of independent spectral and tonal mixtures. This allows sub-word units with different dependences and independent state tying for the two streams, reducing model size and allowing tone-dependent modeling of initials.

We systematically compared the two-stream method with two other approaches that we named *two-model* and *single-stream*. The two-model method yields 5% higher error rates and cannot use one-pass Viterbi decoding, while the single-stream approach requires 30–50% more parameters at similar accuracy.

1. INTRODUCTION

Mandarin Chinese is the most widely spoken language in the world. In contrast to most Western languages it is a *tonal language*. Its five lexical tones constitute a phonemic element of a syllable [2] that is essential for lexical disambiguation [3].

Tones add a further model dimension to speech recognition systems, in addition to the acoustic/phonetic and language model. When adding tone modeling to large-vocabulary continuous-speech recognition (LVCSR), we faced three problems: (1) real-time measurement of pitch (fundamental frequency, F_0), (2) derivation of a suitable pitch feature vector, and (3) integration of tone and acoustic/phonetic model. We treated (1) and (2) in [1]. This paper is about (3).

The approaches to tone integration can be grouped into what we call the *single-stream*, the *two-stream*, and the *two-model* approaches. The two-model approach is an asynchronous method[1]. A non-tonal recognizer determines syllable candidates and segmentation points. Based on these, an independent tone recognizer provides tone scores. Spectral and tonal scores are combined to decide the optimal path [4]. In [5], syllable boundaries are jointly optimized using spectral and tone information. Tone models are independent of the syllable they are applied to [4], and sometimes modeled dependent on their tonal context to integrate coarticulation and context-dependent tone variation (tone sandhi) [5].

The *single-stream* approach as presented by IBM [2] uses a single-stream mixture model and a single augmented feature stream that contains both spectral and pitch features.

The *two-stream* approach, first published in [6], combines the benefits of both, the flexible parameter sharing with one-pass decoding. Multiple mixture streams were also deployed in [3], but stream-specific model dependence was not used.

This paper will first present the three methods within a common framework. We will then discuss the stream-specific model dependence and state tying. Last, we will present experimental results and conclude.

2. A FRAMEWORK FOR RECOGNITION OF MONO-SYLLABIC TONAL LANGUAGES

The goal of a large-vocabulary speech recognizer is, given an acoustic observation $\mathcal{O}$, to find the word sequence $\hat{\mathcal{W}} = (\hat{w}_1 ... \hat{w}_{\hat{N}})$ that most probably generated the observation (the usual maximum-a-posteriori criterion):

$$\begin{aligned}\hat{\mathcal{W}} &= \underset{\mathcal{W}}{\operatorname{argmax}}\, P(\mathcal{W}|\mathcal{O}) \\ &= \underset{\mathcal{W}}{\operatorname{argmax}}\, p(\mathcal{O}|\mathcal{W}) \cdot P(\mathcal{W}) \qquad (1)\end{aligned}$$

In mono-syllabic tonal languages such as Chinese (Mandarin, Cantonese, Taiwanese, etc.) or Thai, every syllable of a word has an associated tone (characteristic pitch contour). Mandarin has a stock of 412 base-syllables and five tones including a neutral tone; there are around 1350 valid combinations. Often, entirely different words share the same phonetic sequence and only differ in tones, such as 媽 (“mā,” first tone, meaning “mother”) and 馬 (“mǎ,” third tone, “horse”).

May $\mathcal{S} = \sigma_1 ... \sigma_M$ denote the sequence of base-syllables, and $\mathcal{T} = \tau_1 ... \tau_M$ the sequence of the associated tones, for the word sequence $\mathcal{W}$. With this, Eq. (1) can be rewritten as:

$$\hat{\mathcal{W}} = \underset{\mathcal{W}\mathcal{S}\mathcal{T}}{\operatorname{argmax}}\, p(\mathcal{O}|\mathcal{S}\mathcal{T}) \cdot P(\mathcal{S}\mathcal{T}|\mathcal{W}) \cdot P(\mathcal{W}) \qquad (2)$$

where $P(\mathcal{W})$ is the usual language model, $P(\mathcal{S}\mathcal{T}|\mathcal{W})$ a pronunciation model, and $p(\mathcal{O}|\mathcal{S}\mathcal{T})$ the usual acoustic model but written in terms of base-syllable and tone sequence.

[1] The two-model approach has conceptual similarities to other asynchronous recognition approaches such as sub-band based systems [7, 8] and asynchronous-transition HMMs [9]. A key difference is that the spectral recognizer delivers rather stable segmentations, which allows for greatly simplified decoder designs.

We now break down $p(\mathcal{O}|\mathcal{S}\mathcal{T})$ to syllable level. May $t_1 \ldots t_M$ denote the times of boundaries between syllables[2], and the m-th syllable span the time interval $T_m = (t_{m-1}, t_m]$. Since in continuous speech the syllable boundaries are unknown, we maximize over all possible syllable-level segmentations:

$$p(\mathcal{O}|\mathcal{S}\mathcal{T}) \quad = \quad \max_{\forall t_1 \ldots t_M} \prod_{m=1}^{M} p(\mathcal{O}_m^{\mathrm{S}} \mathcal{O}_m^{\mathrm{P}} | \sigma_m \tau_m) \qquad (3)$$

$\mathcal{O}_m^{\mathrm{S}} = (o_{t_{m-1}+1}^{\mathrm{S}}, \ldots, o_{t_m}^{\mathrm{S}})$ denotes the sequence of $\underline{s}$pectral features and $\mathcal{O}_m^{\mathrm{P}} = (o_{t_{m-1}+1}^{\mathrm{P}}, \ldots, o_{t_m}^{\mathrm{P}})$ the corresponding $\underline{p}$itch features. The spectral features o_t^{S} are e.g. obtained through MFCC analysis. Robust realtime pitch extraction and the derivation of pitch features o_t^{P} for Mandarin speech recognition has been extensively treated in our previous work [1].

The interesting part of a tonal recognizer for mono-syllabic languages is how $p(\mathcal{O}_m^{\mathrm{S}} \mathcal{O}_m^{\mathrm{P}} | \sigma_m \tau_m)$, the tonal-syllable emission likelihood, is determined. The key question is the degree of statistical dependence between the spectral and tonal model. This manifests in two dimensions:

- *synchrony*: To what degree do we need to force the underlying state segmentations for the spectral and the tonal model to coincide? This question has key influence on decoding architectures.
- *correlation*: To what degree shall the emission distributions of corresponding spectral and tonal states model correlation between spectral and tonal features? This question determines the possibilities of parameter tying.

Depending on these two dimensions, there are three classes of tonal recognizer designs:

- *two-model*: The system enforces only loose synchrony on syllable level[3]; within the syllable, spectral and tonal model are independent, correlation is not modeled.
- *two-stream*: Here, synchrony is enforced at state level, i.e. it is assumed that the state segmentation of both models is identical. Corresponding spectral and tonal emission distributions are assumed statistically independent; state-level correlation is not modeled.
- *single-stream*: Enforces synchrony at state level, and additionally models state-level correlation by augmenting spectral and tonal feature into a single feature stream.

2.1. The two-model approach

This approach assumes asynchronous temporal structure of spectral and tonal features between syllable boundaries. The spectral time-warping path $s_{\sigma_m t}^{\mathrm{S}}$ (base-syllable σ_m) and the tonal path $s_{\tau_m t}^{\mathrm{P}}$ (tone τ_m) are optimized independently. The tonal-syllable emission likelihood is computed as:

$$\begin{aligned} p(\mathcal{O}_m^{\mathrm{S}} \mathcal{O}_m^{\mathrm{P}} | \sigma_m \tau_m) \quad &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4) \\ &\max_{\forall s_{\sigma_m t}^{\mathrm{S}} : t \in T_m} \prod_{t \in T_m} P(s_{\sigma_m t}^{\mathrm{S}} | s_{\sigma_m t-1}^{\mathrm{S}}) \cdot p^{\mathrm{S}}(o_t^{\mathrm{S}} | s_{\sigma_m t}^{\mathrm{S}}) \\ \cdot &\max_{\forall s_{\tau_m t}^{\mathrm{P}} : t \in T_m} \prod_{t \in T_m} P(s_{\tau_m t}^{\mathrm{P}} | s_{\tau_m t-1}^{\mathrm{P}}) \cdot p^{\mathrm{P}}(o_t^{\mathrm{P}} | s_{\tau_m t}^{\mathrm{P}}) \end{aligned}$$

The syllable segmentation is dominated by the spectral model. This allows to simplify the decoding architecture by using only the spectral recognizer to determine syllable boundaries and then applying an independent time-warping engine, constrained by the syllable boundaries, to provide the tone score.

2.2. The two-stream approach

The asynchronous nature of two-model approach above prohibits using standard one-pass Viterbi decoding. If, however, we enforce synchrony on state level, such that there is only one common time-warping path $(s_{\sigma_m \tau_m t})$, then spectral and tonal state emission probabilities are essentially combined in a two-stream manner, and the tonal-syllable emission likelihood is computed as follows:

$$\begin{aligned} p(\mathcal{O}_m^{\mathrm{S}} \mathcal{O}_m^{\mathrm{P}} | \sigma_m \tau_m) \quad &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5) \\ &\max_{\forall s_{\sigma_m \tau_m t} : t \in T_m} \prod_{t \in T_m} P(s_{\sigma_m \tau_m t} | s_{\sigma_m \tau_m t-1}) \\ &\qquad \cdot p^{\mathrm{S}}(o_t^{\mathrm{S}} | s_{\sigma_m \tau_m t}) \cdot p^{\mathrm{P}}(o_t^{\mathrm{P}} | s_{\sigma_m \tau_m t}) \end{aligned}$$

Now, the standard integrated Viterbi search can be used to evaluate Eq. (2), (3), and (5) in a single pass.

2.3. The single-stream approach

The single-stream approach assumes state-level synchrony as well, but uses joint state-emission distributions to model the spectral and tonal features jointly. I.e. correlation is modeled more accurately, but if it is small, parameters may be wasted.

$$\begin{aligned} p(\mathcal{O}_m^{\mathrm{S}} \mathcal{O}_m^{\mathrm{P}} | \sigma_m \tau_m) \quad &= \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (6) \\ &\max_{\forall s_{\sigma_m \tau_m t} : t \in T_m} \prod_{t \in T_m} P(s_{\sigma_m \tau_m t} | s_{\sigma_m \tau_m t-1}) \;\cdot\; p(o_t | s_{\sigma_m \tau_m t}) \\ &\text{with} \quad o_t \quad = \quad \begin{pmatrix} o_t^{\mathrm{S}} \\ o_t^{\mathrm{P}} \end{pmatrix} \end{aligned}$$

3. CONTEXT AND TONE DEPENDENCE

3.1. Stream-specific dependences of sub-word units

The use of context-dependent state-tied sub-word units is a standard technique. Tone models require very different dependences from spectral models. In the two-stream approach, a sub-word unit's states are composed of spectral and tonal states. Thus, the spectral states could, for example, be modeled dependent on the phonetic context, while the tonal states depend on the neighboring tones or the base syllable. If these stream-specific dependences are chosen properly, sub-

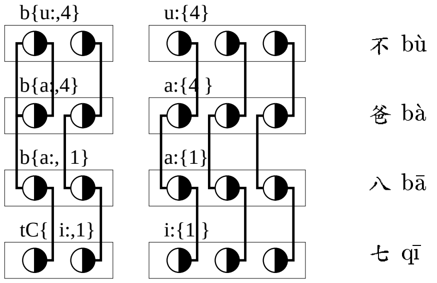

Figure 1: *Illustration of stream-specific model dependence and tying. In this example, sub-word units consist of 2 (preme) or 3 (core-final) states. Circles denote the states, the white half the spectral stream and the black one the tonal stream. The thick lines indicate tying of stream mixtures.*

[2] For simplicity of notation we will completely ignore the existence of silence regions between words and/or syllables.

[3] Some systems synchronize at half-syllable boundaries. According to our experiments, the accuracy difference is minimal.

Table 3: *Comparison of various choices of tone dependences.*

id	tone dependent?				MAT			PCD			863			av.	av.
	preme		core-final		size	TER	CER	size	TER	CER	size	TER	CER	size	CER
	$\mathcal{O}^{\mathrm{S}}$	$\mathcal{O}^{\mathrm{P}}$	$\mathcal{O}^{\mathrm{S}}$	$\mathcal{O}^{\mathrm{P}}$	[M]	[%]	[%]	[M]	[%]	[%]	[M]	[%]	[%]	change	change
no tonal features used															
B1	—	n/a	—	n/a	3.0	n/a	22.8	1.8	n/a	16.9	2.5	n/a	10.2	ref	ref
B2	—	n/a	√	n/a	4.6	47.1	23.1	3.0	47.5	15.9	3.4	41.2	8.0	+52.0%	−8.7%
two-stream approach															
T1	—	—	—	√	3.0	26.4	17.6	2.0	22.2	11.5	2.7	24.9	6.1	+6.4%	−31.7%
T2	—	—	√	√	4.6	26.0	16.9	3.1	22.4	11.9	3.6	22.9	6.4	+56.5%	−30.9%
T3	—	√	—	√	3.1	24.0	16.0	2.0	19.8	11.0	2.7	21.9	6.4	+7.5%	−34.0%
single-stream tonal feature vectors															
S1	—		√		4.5	25.2	16.3	3.1	20.5	11.4	3.9	23.0	6.6	+59%	−32.1%
S2	√		√		8.9	17.9	21.9	5.1	15.1	11.3	7.2	19.7	6.6	+189%	−24.1%

word units with all desired combinations of phonetic and tonal dependence can be synthesized by combining the proper mixtures [6]. Furthermore, data-driven state tying methods can be applied differently for the two streams[4]. While for phonetic-context dependent models tying is often constrained to corresponding allophone states, tying of tone models would be constrained to the same tone. The concept is illustrated in Figure 1.

The single-stream approach does not allow this. Instead, each unit needs to have the required context and tone dependence at the same time. Due to limited training material, only one dependence at a time is feasible in many cases.

3.2. Tonal modeling of initials

[2] suggests that the tone information is concentrated in a syllable's main vowel. However, we found that in addition to that, some initials (particularly liquids and nasals) carry part of the tone contour as well, and that some accuracy gain can be achieved by tone-dependent modeling of initials (premes).

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

The Philips Mandarin LVCSR research system is HMM-based. We use the usual MFCC features, tone features as described in [1], right-context dependent preme/core-final models [10], and, unless otherwise noted, 32-Gaussian mixtures.

Table 1: *Corpus characteristics.*

	MAT-2000		PC Dictation		863	
	Train	Test	Train	Test	Train	Test
Type	Telephone		Microphone		Microphone	
Accent	Taiwan		Taiwan		Mainland	
#Speakers	2232	22	241	20	2×83	2×6
#Utterances	113475	500	27606	199	92948	2×240
#Syl./Utt.	4.5	9.5	30.1	35.5	12.1	12.6
Lexicon size	—	56303	—	42038	—	56064
CPP_{tri}	—	56.7	—	51.1	—	41.3

Experiments were conducted on three Mandarin continuous-speech corpora: MAT-2000 (telephone, Taiwan, [11]), a PC dictation database (PCD, microphone, Taiwan), and the database of the Chinese 1998 National Performance Assessment (National Hi-Tech Project 863, microphone dictations [12]). The test set for MAT is the official development test set of the 2000 Taiwan benchmarking (see also [13]). Recognition for MAT and PCD is speaker-independent, while for 863 it is gender-dependent and the gender is known during decoding.

Table 1 summarizes the corpora and also shows lexicon size and character-normalized test-set perplexities of the trigram language model (CPP_{tri}). For the 863 "System Performance Test," the standard language-model training corpus includes the test set. To obtain realistic performance figures, we "cleaned" this corpus by carefully removing all 480 test sentences (after which perplexity increased from 14.4 to 41.3).

4.2. Comparison of basic structures

We compared the improvement of character error rates (CER) from tone modeling for the three approaches. Table 2 shows CERs for all three corpora and the relative gains over the non-tonal baseline, averaged over the corpora (right column).

Table 2: *Comparison of different tone-modeling approaches (LVCSR character error rates for the respective best setups).*

id	tone-modeling	database			av. rel.
	approach	MAT	PCD	863	change
B1	non-tonal baseline	22.8%	16.9%	10.2%	ref
M1	two-model	16.9%	12.0%	**6.4**%	−30.7%
T3	two-stream	**16.0**%	**11.0**%	**6.4**%	−**34.0**%
S1	single-stream	16.3%	11.4%	6.6%	−32.1%

Experiment B1 is the non-tonal baseline. M1 denotes the two-model approach, T3 the two-stream setup, and S1 the single-stream method. In this comparison, the respective best setups are used; M1 and T3 use tonal premes, while the model in S1 uses toneless premes. Of the three approaches, the two-stream method leads by a small margin.

4.3. Choice of tone dependence

We experimented with various choices of tone dependences of involved models, as shown in Table 3. In addition to character error rates, this table also shows model size and tone error rates (TER, see box). B1 is the non-tonal baseline.

[4] indicates that spectral features carry tone information to some degree, such that using tone-dependent models even without pitch-related features should yield part of the gain. Experiment B2 shows that this effect is small but indeed present: an average 9% CER reduction is achieved. In particular, for the very homogeneous 863 data, the gain is 21.6%.

Experiment T1 is the two-stream setup, where tone-dependent models are trained only for the core-finals, while the premes

[4] The idea of independent state tying for different streams was introduced by Sagayama in the context of sequential asynchronous-transition HMMs. State tying was applied across time independently for the cepstral and the delta stream [9].

Tone Error Rate (TER)

We define the TER as the number of tonal syllables that are recognized correctly, while *knowing* the true non-tonal syllable sequence. We take into account that the optimal syllable segmentation may depend on the tones. The TER is computed as follows:

1. convert each sentence of the test set into a finite-state network as follows:
    - form a sequence of sub-networks, one sub-network for each character in the sentence;
    - each character's sub-network contains all possible tonal versions of the character's base-syllable as parallel branches;

    I.e., one path through the network constitutes the correct sentence prununciation (sequence of tonal syllables), all other paths differ only in the tones;
2. recognize each sentence according to its network;
3. count errors. The number of substitutions is the TER (insertion / deletion errors are impossible).

Note that in [1], we used a simplified method.

are tone-independent (pooled over the tones). The spectral model is from B1. The average CER improvement is 31.7% at a small overall parameter increase of 6.4%. In T2, we combined the tone model with B2; the spectral model increases by around 50%, but we observe no CER gain. In T3, we also modeled the premes tone-dependent. While TERs improve between 9 and 12%, a CER gain is present but modest (3.4% on average; 9.1% on MAT).

S1 is the standard single-stream setup, similar to [2]. Good CERs between T1 and T3 are achieved, but at a 59% increase of model parameters. In S2, we also tried tone-dependent modeling of the premes. However, this leads to a huge increase of model parameters (nearly factor of 3), since premes are now context *and* tone dependent at the same time. Although tone error rates are vastly improved, discriminability of base syllables suffers. Neither change of tying thresholds nor less mixture splitting could improve that. We believe that we just wasted a lot of parameters to model the weak tone dependence of the spectral features, and that some units now have even too few observations for reliable state tying.

4.4. Comparison of model size

In Figure 2, we compare the two-stream and single-stream methods for different model sizes, since accuracy strongly depends on the number of model parameters (two-model results are not shown). Aside of state tying, no size-reduction methods were applied here. For all three databases, the best two-stream setup outperforms the best single-stream setup w.r.t. accuracy and model size.

5. CONCLUSION

We described the two-stream approach to tone recognition for Mandarin LVCSR and compared it to the two-model and single-stream approach. By tone modeling, relative reductions of character error rates of 32-34%, averaged over three corpora, have been achieved, the largest gains ever reported.

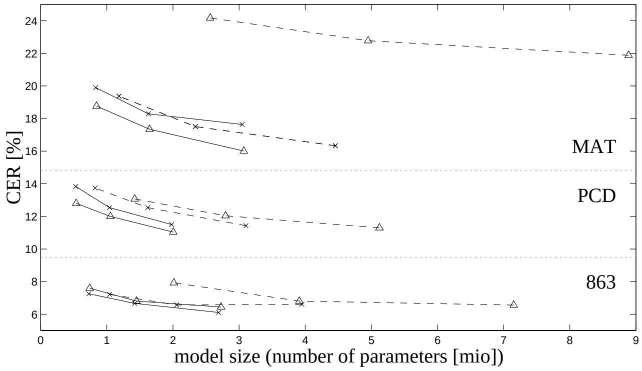

***Figure 2**: CER as a function of model size (using 8, 16, or 32 mixture components). Solid lines denote two-stream results; dashed lines single-stream. Triangles denote results for tone-dependent premes, crosses for tone-independent premes.*

The benefit of two-stream modeling for tone recognition is that the spectral and tonal mixtures of a sub-word unit can be modeled with different dependence, and that state tying can be applied in different dimensions according to different criteria. The flexible parameter tying makes it feasible to apply tone modeling to the *premes* as well, which yielded up to 9% performance gain, 3.4% averaged over corpora.

Overall, the two-stream method slightly outperforms the two other tone-modeling approaches in terms of accuracy (3-5% averaged over three corpora). Compared to the single-stream approach, model size is significantly reduced by a third. As opposed to the two-model approach, the standard one-pass Viterbi decoding is applicable.

6. REFERENCES

[1] H. Huang and F. Seide. Pitch tracking and tone features for Mandarin speech recognition. In *Proc. ICASSP'2000*, Istanbul, 2000.

[2] C.J. Chen et al. New methods in continuous Mandarin speech recognition. In *Proc. EUROSPEECH*, Rhodos, 1997.

[3] S. Liu et al. The effect of fundamental frequency on Mandarin speech recognition. In *Proc. ICSLP'98*, Vol. 6, pp. 2647–2650, Sydney, 1998.

[4] C.H. Lin et al. Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units. In *Speech Communication*, 18 (1996), Elsevier Science B.V.

[5] H.M. Wang et al. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data. In *Proc. ICASSP'95*, pp. 61–64, Detroit, 1995.

[6] T.H. Ho et al. Phonetic state tied-mixture tone modeling for large vocabulary continuous Mandarin speech recognition. In *Proc. EUROSPEECH'99*, pp. 883–886, Budapest, 1999.

[7] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. ICSLP'96*, pp. 426–429, Philadelphia, 1996.

[8] S. Tibrewala and H. Hermansky. Sub-band recognition of noisy speech. In *Proc. ICASSP'97*, pp. 1255–1258, Munich, 1997.

[9] S. Sagayama et al. Asynchronous-transition HMM for acoustic modeling. In *Proc. ASRU'99*, pp. 99–102, Colorado, 1999.

[10] F. Seide and N. Wang. Phonetic modelling in the Philips Chinese continuous-speech recognition system. In *Proc. ISCSLP'98*, pp. 54–59, Singapore, 1998.

[11] H.C. Wang et al. MAT-2000 – Design, collection, and validation of a Mandarin 2000-speaker telephone speech database. In *Proc. ICSLP'2000*, Beijing, 2000.

[12] R.H. Wang. National performance assessment of speech recognition systems for Chinese. In Proc. Oriental COCOSDA Workshop '99, pp. 41–44, Taipei, 1999.

[13] L. Liao et al. Improvements of the Philips 2000 Taiwan Mandarin benchmark system. In *Proc. ICSLP'2000*, Beijing, 2000.