Phonetic Transcription Verification with Generalized Posterior Probability

Lijuan WANG[1], Yong ZHAO[2], Min CHU[2], Frank K. SOONG[2], Zhigang CAO[1]

[1]Department of Electronic Engineering, Tsinghua Univ. China
[2]Microsoft Research Asia, Beijing, China
wlj01@mails.tsinghua.edu.cn
{yzhao, minchu, frankkps}@microsoft.com

Abstract

Accurate phonetic transcription is critical to high quality concatenation based text-to-speech synthesis. In this paper, we propose to use generalized syllable posterior probability (GSPP) as a statistical confidence measure to verify errors in phonetic transcriptions, such as reading errors, inadequate alternatives of pronunciations in the lexicon, letter-to-sound errors in transcribing out-of-vocabulary words, idiosyncratic pronunciations, etc. in a TTS speech database. GSPP is computed based upon a syllable graph generated by a recognition decoder. Testing on two data sets, the proposed GSPP is shown to be effective in locating phonetic transcription errors. Equal error rates (EERs) of 8.2% and 8.4%, are obtained on two testing sets, respectively. It is also found that the GSPP verification performance is fairly stable over a wide range around the optimal value of acoustic model exponential weight used in computing GSPP.

1. Introduction

Large speech corpora have become standard tools for speech research and product/service development. However, before the corpora can be used for their designated purposes, they often need to be manually checked, annotated or segmented. Phonetically transcribed databases have long been used in linguistic research, both for explorative and hypothesis testing purposes. More recently, they have been shown to be useful for developing automatic speech recognition and synthesis systems.

Take the concatenation-based Text-to-speech (TTS) synthesis as an example [1]. When a sequence of phonemes is to be synthesized, proper acoustic units (typically phones, diphones or units of non-uniform length) are selected from a pre-recorded corpus. The synthesis tokens are selected from the corpus, according to their phonetic labels and contexts, etc. Therefore, accurate phonetic transcription is critical to the final quality of synthesized speech.

As it has been well known, the production of manual labeling and transcriptions is time-consuming, costly and error-prone. Recourse to automatic or semiautomatic annotation of speech data is therefore desirable.

To obtain phonetic labeling of speech automatically is to do forced recognition (or forced alignment) [5][6]. Rather than recognizing speech as a string of unknown words in automatic speech recognition (ASR), the word strings (or the orthographic transcription of the utterance) are given beforehand. Given the orthographic transcriptions, forced recognition can check word for word to see which may be mispronounced. In order to do so, we may need to provide as many possible pronunciations for each word.

The automatic forced recognition method cannot guarantee the error-free phonetic transcription because of the following reasons:

- Reading errors, or orthographic pronunciation errors;
- Incomplete list of all possible pronunciations in the lexicon, including letter-to-sound errors for out-of-vocabulary words;
- Idiosyncratic pronunciations of a speaker.

In this paper, we propose a confidence measure for verifying phonetic transcriptions with *generalized syllable posterior probability* (GSPP). Laborious human checking effort can be alleviated or even eliminated and good-quality phonetic transcriptions of large amounts of speech material can be obtained.

2. Generalized posterior probability

Generalized posterior probability (GPP) is a probabilistic confidence measure for verifying the recognized (hypothesized) entities at different levels, e.g., subword, word and utterance level [2][3][4]. It was first applied to verification at the word level under various testing conditions. GWPP assesses the reliability of a focused word by "counting" its reappearances in the word graph and reweighting the corresponding path likelihood exponentially both acoustically and language model wise and normalized by the acoustic probability. GWPP is defined as

$$p([w;s,t]\,|\,x_1^T) = \sum_{\substack{M,[w;s,t]_1^M \\ \exists n,\ 1\le n\le M \\ w=w_n \\ |s_n-s|\le\Delta,|t_n-t|\le\Delta}} \frac{\prod_{m=1}^{M} p^{\alpha}\left(x_{s_m}^{t_m}\,|\,w_m\right)\cdot p^{\beta}\left(w_m\,|\,w_1^M\right)}{p\left(x_1^T\right)} \quad (1)$$

where $[w;\ s,\ t]$ is the focused word w with its starting time s and ending time t, x_1^T is the whole sequence of acoustic observations, M is the number of words of a string in the graph, α and β are the exponential weights for the acoustic and language models, respectively. $P(x_1^T)$, the acoustic

[1] This work has been done when the first author is a visiting student in Microsoft Research Asia.

probability of all observations, can be computed by summing the likelihood of all paths in a given search space, e.g., word graph.

GWPP has been demonstrated to deliver robust performance on word verification at different search beam widths [2], signal-to-noise ratios, etc.

3. Generalized syllable posterior probability

The computation of GWPP, including the acoustic observation probability and P(x_1^T), is carried out in a reduced search space (e.g., word graph or N-best list). These graphs are typically constructed by using pronunciations in the lexicon with a word-level language model constraint such as word n-grams. This framework has been shown to be effective in computing GWPP by combining multiple knowledge sources in an integrated search space [2]. However, in our specific TTS application the use of the word as the main unit of representation experiences some difficulties.

One main problem is that it is rather difficult to detect partial and minor pronunciation variations of a word, especially long, in continuous speech. In ASR word pronunciation variations only causes a second order effect in decoding, in TTS applications such variations need to be resolved with higher resolution. Left-to-right word fitting algorithms reveal that sometimes the hypothesized word still dominates the word graph, even if certain canonical phoneme in a word pronunciation has not been correctly articulated. Meanwhile, constrained by the word lexicon and N-gram language model, partial hypothesis candidate with exactly matched phoneme sequence might be discarded before completion. A similar phenomenon is the errors existed in the pronunciation lexicon, especially for the out-of-vocabulary words, the pronunciations of which are generated by statistical letter-to-sound conversion algorithm [7].

Another problem is the out-of-vocabulary (OOV) words. No matter how large the vocabulary is, OOV words are almost unavoidable since the vocabulary of an active language is changing continuously.

To solve the above problems, we propose to use subwords as the basic units. Or the complete set of basic units of a given language like phonemes and syllables. The phonemes form the most efficient set but not as reliable as syllables for tangible decoding due to their shorter durations and no guarantee of vowel nuclei. We can define a complete set of syllables in a language and generate a network of all the possible syllable sequences. This network is complete because it represents all possible syllable sequences that can be spoken. Furthermore, it introduces more constraints in the search space than a pure network of phonemes. It is also possible to detect hesitations, corrections, and other spontaneous speech phenomena and to detect alternate pronunciation by the speaker.

Therefore we generalize the posterior probability from word to syllable. The equation of *generalized syllable posterior probability* (GSPP) is defined as

$$p([syl;s,t]\,|\,x_1^T) = \sum_{\substack{M,[syl;s,t]_1^M \\ \exists n,1\le n\le M \\ syl=syl_n \\ |s_n-s|\le\Delta,|t_n-t|\le\Delta}} \frac{\prod_{m=1}^{M} p^{\alpha}\left(x_{s_m}^{t_m}\,|\,syl_m\right)}{p\left(x_1^T\right)} \qquad (2)$$

where [*syl*; *s*, *t*] is the focused syllable with its starting time *s* and ending time *t*, x_1^T is the sequence of acoustic observations, *M* is the number of syllables of a path in the syllable graph, α is the exponential weights for the acoustic models. Since GSPP concentrates more on acoustic details than language model (e.g., word content), only the statistical acoustic confidence is considered for the hypothesized syllable.

Similar to GPP [2], three issues, the reduced search space, time relaxation registration, and re-weighted acoustic model likelihood are employed in computing GSPP.

A syllable graph, rather than a word graph, is served as the reduced search space for GSPP. A syllable unigram is used as language model to generate the syllable graphs of rich acoustic candidate hypotheses.

We assign zero language model weight to all hypotheses in the search space, and the values of language model likelihood are ignored. The acoustic likelihoods reweighting is to prevent the syllable posterior probability from being dominated by just a few top strings with high likelihoods, and to accommodate the modeling discrepancies in the practical implementations, including:

- Unbounded dynamic range: In an acoustic model of Gaussian mixture, acoustic likelihoods obtained from pdf have an unbounded dynamic range.
- Likelihood computation frequency: Acoustic likelihoods are computed every frame.
- Independence assumption: Neighboring acoustic observations are assumed to be statistically independent.
- Reduced search space: The full search space is pruned to a syllable graph.

4. Experiment setup

4.1. Data preparation

The speech corpus for our experiments is English Corpus we use for constructing Microsoft TTS system. It is a large vocabulary, continuous, read speech corpus recorded by a professional female speaker, containing 6,500 utterances in total and covering various phonetic contexts. The phonetic transcription of this corpus is manually annotated and verified by several transcribers. In particular, two testing sets of 500 sentences each, denoted as set1 and set2, are used to evaluate GSPP.

4.2. Syllable based recognition

The whole corpus is used to train the speaker dependent acoustic HMMs. 39 acoustic features (12 MFCC + 12ΔMFCC + 12ΔΔMFCC + logE +ΔlogE +ΔΔlogE) are used. 4 Gaussian components per mixture are used for modeling the output probability density function of each tri-phone tied state. A lexicon of 7,800 syllables and a syllable unigram language model are used to generate the wide beam syllable graphs.

4.3. Evaluation

To evaluate the proposed GSPP, two test sets are created. The first is the positive set where the expected syllable is spoken, and the instances are all expected to be accepted. This is known as the 'accept set.' It is relative tricky to come up with a negative set, where all the instances should be rejected. We decide to create an artificial set. In the 'reject set' an erroneous transcription $\mathcal{S}'$ is created for each testing token $\mathcal{S}$ (a *correct* given syllable). What we are more interested in is the cases where people misread the correct transcription by uttering similar (therefore confusable) words. We deliberately choose an $\mathcal{S}'$ confusable to syllable $\mathcal{S}$. The substitution by $\mathcal{S}'$ is allowed if the following requirements are met.

$\mathcal{S}$ and $\mathcal{S}'$ are equal in length (number of phonemes) but different by one phoneme.

The absolute difference between their logarithmic language model scores in syllable unigram is below a preset threshold (0.3 in this experiment).

Table 1: An example of a phone replacement

transcription	reply		to	his	tailor	
syllable spoken ($\mathcal{S}$)	r ih	p l ay	t uw	h ih z	t ey	l ax r
accept set ($\mathcal{S}'$)	r ih	p l ay	t uw	h ih z	t ey	l ax r
reject set ($\mathcal{S}'$)	***s*** ih ***d*** ih r ***iy***	p l ***ax*** ***k*** l ax p l ***oy*** p l ***ae*** p l ***ey*** p ***r*** ay p l ***iy***	t ***iy*** ***y*** uw t ***ax*** t ***ih***	***jh*** ih z h ***ae*** z ***w*** ih z h ih ***m*** ***ch*** ih z ***b*** ih z ***z*** ih z	***v*** ey ***k*** ey ***b*** ey t ***ay*** ***p*** ey ***s*** ey ***dh*** ey …	***dh*** ax r ***ch*** ax r ***p*** ax r ***n*** ax r ***m*** ax r ***jh*** ax r ***f*** ax r …

Table 2: Summary of the two testing sets

	set1	set2
# utterances	500	500
# words	3,848	4375
# syllables in accept set	5,394	7,257
# syllables in reject set	42,564	59,494

Table 1 shows examples from these two sets. Actually in the 'reject set', a syllable substitution is implemented as one phoneme substitution by looking through the syllable lexicon for all possible substitutes. For each time, only one correct syllable is tagged as incorrect in a sentence. For the two testing set, namely set1 and set2, both a 'accept set' and a 'reject set' are derived, as shown in Table 2. The boundaries for the hypothesized syllable are derived from phoneme boundary obtained by forced alignment [7]. GSPP for each hypothesized syllable in the 'accept set' and the 'reject set' is then computed.

The performance of GSPP is represented by its Receiver Operating Characteristic (ROC). The ROC curve is the plot of a false rejection rate (calculated on the 'accept set' by Eqn.3) with respect to a false acceptance rate (calculated on the reject set by Eqn.4) at each threshold value. These curves are also used to determine the optimal acoustic model weights and rejection threshold.

Another criterion is the equal-error-rate (EER). The EER is obtained by adjusting the threshold value so that the false acceptance and the false rejection error rate are equal.

$$\text{false reject rate} = \frac{\text{\#of false rejections}}{\text{total \#of hypothesized syllables}} \times 100\% \qquad (3)$$

$$\text{false accept rate} = \frac{\text{\#of false acceptances}}{\text{total \#of hypothesized syllables}} \times 100\% \qquad (4)$$

5. Experimental results

In order to find the optimal acoustic model weight and the corresponding rejection threshold, a full grid search is applied to set1. The ROC curves at different exponential weights of acoustic model (varied from 0.0 to 10.0, the curve color is changed from blue to red) are plotted in fig. 1. The results show that when the acoustic model weight α is below 0.03, the GSPP performance degrades as α decreases. When α increases from 0.03 to 10.0, the ROC curves are almost on top of each other. From another point of view, fig. 2 shows the EER of GSPP saturates when α exceeds 0.03. The above results show that the verification efficiency of GSPP is robust to the change of α. Within a rather broad range of α, say from 0.03 to 10, the EER of GSPP stays at 8.2%.

In the GSPP framework, the introduction of the acoustic model weights offers further control on the relative importance of the ranked hypotheses. For larger weights, more emphasis put on the higher ranked hypotheses. Smaller weights, on the other hand, take more hypotheses into consideration in computing GSPP. In the extreme case, when the weights are set to infinity, only the best hypothesis is considered. In our experiments, high ranked hypotheses become dominant components in GSPP calculation when the acoustic model weights α reaches 0.03. When α is increased further, the higher ranked hypotheses is enhanced and thus the performance gets steady. This result is in some sense in disagreement with that in ASR using GWPP: smaller acoustic model weights always yield better verification performance. The reason may lie on that the TTS corpus used in our experiments is a single speaker, read style corpus, which is more homogenous in speaking rate, pronunciation, etc., than the speaker-independent, sometimes noisy ASR. Since the speaker-trained model is sharp, the higher ranked hypotheses in the syllable graph are more reliable. Fig. 2 also shows the EER saturates at 8.4% on another testing set 'set2'.

The results on the two testing sets show the consistency and efficacy of the verification performance of GSPP.

6. Analysis and discussion

6.1. Inadequate HMM discrimination

In our experiments, 384 types of phoneme substitutions appear in the 'reject set' derived from the two testing sets. The verification performance of GSPP is not uniform across all phoneme substitutions. The EER for detecting certain substitutions exceeds 50%, such as /er/--/ax/, /ah/--/aa/, /uh/--/ih/, /iy/--/ih/, /t/--/d/, /z/--/s/, etc., although the average EER is 8.2%. It shows that the confusion between the phonemes in the same place of articulation [7], like /t/--/d/, or with the similar features: +/- high, +/- low, +/- front, +/- back, and +/- round, like /iy/--/ih/, are the most difficult for GSPP to verify. The reason is that GSPP is computed based on a syllable graph, a byproduct of LVCSR decoding. But due to the inadequate discrimination of acoustic HMM, the models can not differentiate certain phonemes, similarly for GSPP.

6.2. GSPP overestimation

For two phonetically identical or similar syllables adjacent to each other in the transcription, their corresponding GSPPs might be overestimated, due to the relaxed time registration in computing GSPP. Thus, verifications of both syllables might be unreliable. It can happen between two adjacent words, like in 'his history' /h ih s h ih s t r iy/, or within a word like 'sissy' /s ih s ih/. Although such instances occur not so frequently, proper verification still deserves some special attention.

7. Conclusion

GSPP is proposed as a reliable confidence measure to verify phonetic transcriptions of a speech corpus. A syllable graph serves as the reduced search space in computing GSPP. It is shown that GSPP yields an EER of 8.2% and 8.4% on the two testing sets, respectively. It is also found that the phonetic verification performance of GSPP is fairly stable over a wide range around the optimal exponential acoustic weight for computing GSPP.

8. References

[1] Fackrell, J., Skut, W., and Hammervold, K. "Improving the accuracy of pronunciation prediction for unit selection TTS," in *Proc. EUROSPEECH-2003, pp. 2473-2476, Geneva, Switzerland, September 2003.*

[2] Soong, F.K., Lo, W.K., and Nakamura, S. "Optimal acoustic and language model weights for minimizing word verification errors," in *Proc. ICSLP-2004, Jeju, October 2004.*

[3] Soong, F.K., Lo, W.K., and Nakamura, S. "Generalized word posterior probability (GWPP) for measuring reliability of recognized words," in *Proc. SWIM-2004, Hawaii, January 2004.*

[4] Wessel, F., Schluter, R., Macherey, K., and Ney, H., "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Proc., Vol. 9, pp.288-298, 2001.*

[5] Binnenpoorte, D. and Cucchiarini, C., "Phonetic transcription of large speech corpora: How to boost efficiency without affecting quality," in *Proc. ICPhS-2003, 2003.*

[6] Cucchiarini, C., Binnenpoorte, D., and Goddijn, S., "Phonetic transcriptions in the Spoken Dutch Corpus: how to combine efficiency and good transcription quality," in *Proc. EUROSPEECH-2001, pp. 1679–1682, 2001.*

[7] Huang, X.D., Acero, A., and Hon, H., Spoken Language Processing, *Prentice Hall, New Jersey, 2001.*

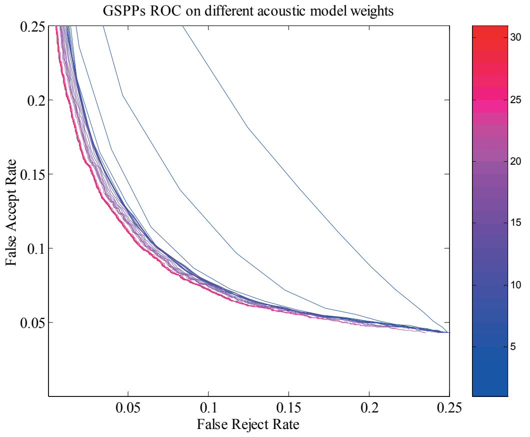

Figure 1: GSPP's ROC with different acoustic model weights on set1.

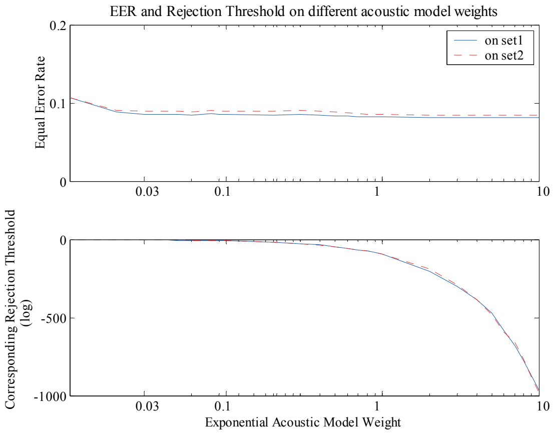

Figure 2: EER and corresponding rejection threshold with different acoustic model weights on set1 and set2.