Bilateral G2P Accuracy: Measuring the effect of variants

Oluwapelumi Giwa[1,2] and Marelie H. Davel[1,2]
[1]Multilingual Speech Technologies, North-West University, South Africa.
[2]CAIR, CSIR Meraka, South Africa.
oluwapelumi.giwa@gmail.com, marelie.davel@gmail.com

Abstract— **Incorporating pronunciation variants in a dictionary is controversial, as this can be either advantageous or detrimental for a speech recognition system. Grapheme-to-phoneme (G2P) accuracy can help guide this decision, but calculating the G2P accuracy of variant-based dictionaries is not fully straightforward. We propose a variant matching technique to measure G2P accuracy in a principled way, when both the reference and hypothesised dictionaries may include variants. We use the new measure to evaluate G2P accuracy and speech recognition performance of systems developed with an existing set of dictionaries, and observe a better correlation between G2P accuracy and speech recognition performance, than when utilising alternative metrics.**

I. Introduction

The pronunciation model is one of the key components of an automatic speech recognition (ASR) system and is most often modelled as an explicit pronunciation dictionary (a list of words and their predicted pronunciations). The dictionary is typically developed manually, using human expertise, or created in a data-driven manner from a training sample. Capturing acoustic variabilities, such as dialect, differences in semantics or accents, could result in incorporating different pronunciation variants into the dictionary [1]. Capturing pronunciation variation in a dictionary is controversial as, in practice, it can either increase or decrease ASR performance.

Prior to developing an ASR system, G2P accuracy can be measured to help guide the choice of pronunciation dictionary. Ambiguity arises when standard G2P accuracy measures are used to score variants; that is, measuring predicted pronunciation variants against the gold standard reference dictionary that also includes variants. In some of our earlier work, little correlation was found between G2P accuracy and ASR performance in such cases [2]. One apparent reason for artificially improved G2P accuracy, depending on measure selected, is that either over-generation or under-generation of variants is unfairly advantaged.

In this work, we propose balancing these two errors in a more principled way and developing a straightforward but effective technique for measuring G2P accuracy that correlates well with observed ASR performance. We experiment with systems developed in four South African languages, namely Afrikaans, English, Sesotho and isiZulu.

This paper is structured as follows: Section II provides background on pronunciation variation, G2P accuracy measures and issues with pronunciation variants. Section III provides an overview of our approach to measuring G2P accuracy, Section IV describes the experimental design, and Section V contains the results of the analysis. Section VI summarises our findings.

II. Background

ASR systems trained on data from one dialect often do not generalise to another dialect, due to the different acoustic realisation of words [3]. This mismatch is represented by two ASR components: the acoustic and pronunciation models. Addressing this issue through acoustic model adaptation is usually preferred over incorporating pronunciation variants in the dictionary, and the latter technique used for including truly distinct pronunciations of words, for example, as observed for proper names or code-switched words (words of a different language embedded in speech).

Adding pronunciation variants to a dictionary is controversial, as it can be considered advantageous or detrimental to the system. Brunet and Murthy [4] investigated the effect of pronunciation variation on recognition accuracy. In their work, a data-driven approach was employed to observe pronunciation variation at syllable level. For each word, a syllable-level pronunciation was obtained using the CMU pronunciation lexicon[1] and NIST syllabification software[2]. Their experiment was performed on the TIMIT corpus[3] of read speech. They observed that the inclusion of prominent variants improve recognition accuracy compared to when rare variants are incorporated. In related work, Hahn et al. [7] investigated the influence of pronunciation variants on recognition accuracy. In their work, up to three variants could be incorporated into the pronunciation lexicon. They observed that adding variants were helpful to some languages, while harmful to others. Previous work shows that pronunciation variants vary significantly across domain and dialect. Jouvet et al. [8] evaluated the average number of pronunciation variants per word that is required to optimise recognition accuracy in the context of French broadcast news data. Given a range of thresholds, pronunciation variants were selected

[1]An open source pronunciation dictionary created for speech recognition and synthesis research. It provides a mapping from orthographic to phonetic based on the ARPAbet phoneme set [5].

[2]Software that marks syllable boundaries and its implementation is based on Kahn's theory of syllabification [6].

[3]An acoustic-phonetic continuous speech designed for the development and evaluation of automatic speech recognition systems

based on their posterior probability. In this specific example, an average of 1.07 to 1.66 variants were produced per word, and the optimal result was quite low (1.1 average per word, or less).

When measuring general G2P accuracy, standard measures are based on the Levenshtein distance, measured either at the word or phoneme level. Word error rate (WER) and phoneme error rate (PER) are both widely used, with PER calculated as the total number of insertions, substitutions and deletions required to transform the hypothesised to reference pronunciation (the Levenshtein distance) divided by the total number of phonemes in the reference pronunciation. Hixon et al. [9] used these metrics together with a proposed weighing based on the probability of confusion of two phonemes. In their work, a weighted phoneme substitution matrix can better score the confusability of the pronunciations with respect to the reference pronunciations.

For variant-based dictionaries, that is, when measuring predicted pronunciation variants against the gold standard reference dictionary that also includes variants, it is not fully straightforward how variants should be accounted for. Each reference variant can be treated as a separate word, or as in [1], the variants linked to their producing word. The two alternative measures used here were referred to as single-best and variant-based versions of WER and PER.

For a given word, single-best variant accuracy is computed by obtaining the single best-matching reference-hypothesis pair based on their phone accuracy score. Variant-based accuracy is computed by obtaining the best-matching hypothesised pronunciation for each reference pronunciation. This variant-based G2P metric is computed per word, by evaluating each reference pronunciation against all hypothesised pronunciations (for that specific word) to obtain the best-matching pronunciation, as well as the accuracy score for that reference-hypothesis pair. These scores are then averaged across all pronunciation variants of the specific word occurring in the reference dictionary. Variant-based WER can therefore only be 0% if, for any given word, there is a hypothesised pronunciation that matches the reference pronunciation for every single pronunciation variant occurring in the reference dictionary. An additional measure used, Matching Variant Percentage (MVP) provides an indication of the extent to which variants were under- or over-generated. It is calculated as the ratio between the average number of required variants to the average number of obtained variants.

III. Approach

In earlier work [2], we observed that G2P accuracy and ASR performance do not always correlate well if pronunciation variants occur. Also, we noticed that the way pronunciation variants are dealt with when measuring accuracy has a significant effect on the result. We address this by using *bilateral scoring*, an approach described in more detail in Section III-A below.

A. Bilateral scoring

Bilateral scoring helps obtain a better indication of the balance between accuracy and number of variants by considering all the variants in both the reference and hypothesised lexicons. Specifically, it first pairs up all variants in the reference dictionary and all variants in the hypothesised dictionary, one-by-one. Per word, if there are more reference variants than hypothesised variants, each of the hypothesised variants will first be mapped to its best matching reference variant, and then hypothesised variants will be 're-used' to form pairs with all matching reference variants, and vice versa. Algorithm 1 describes this process. The specific score used to determine the distance between the two pronunciations – the phone-based dynamic programming (PDP) score – is described in the next section.

See Table I for a demonstration of the difference between unilateral and bilateral scoring.

B. Phone-based dynamic programming

In order to map the hypothesised variant with its best reference variant, a string alignment is used to align the observed and reference string with each other to produce an alignment score. This score can be calculated using any algorithm that gives an indication of cost or distance. Specifically, for the current task, we use phone-based dynamic programming [10] (PDP) to obtain the alignment score.

The PDP scoring algorithm is designed to provide a rank ordering between two given pronunciation strings. Alignment with larger scores is more likely to match the reference string while lower scores indicate otherwise. To obtain a PDP score, standard dynamic programming with a scoring matrix is used to map reference and observed phone strings with each other. In practice, dynamic programming is used as a measure of similarity between the reference and observed string. The scoring matrix is used to penalise specific substitutions between phones in the observed and reference string.

A scoring matrix can either be obtained manually or derived from the data (data-driven). A flat or uniform scoring matrix can be generated manually [10], associating a specific cost with each phone substitution, deletion or insertion. For example, a similarity score of +1 can be assigned if the symbols are identical, -1 if symbols differ and -0.5 for a gap either in the reference or observed.

A data-driven scoring matrix, also known as a weighted scoring matrix, is derived directly from the data. A flat scoring matrix is used for initialising the scoring matrix. Then, by aligning all the strings, a count matrix is computed that contains the number of times a reference phone was recognised as an observed phone. A further smoothing process of the matrix is carried out by adding a single count to each matrix entry (Laplace smoothing). A set of log posterior probability is calculated from the smoothed matrix. To obtain a set of optimal log likelihoods per reference phoneme, values are summed across intra-word phones and the overall deletion probability added. More detail is presented in [10]. The weighted scoring is therefore analogous to Hixon's approach, described in Section II.

TABLE I

EXAMPLE: COMPARING UNILATERAL AND BILATERAL V-PA FOR HYPOTHETICAL WORDS 'ABUSE', 'APE', 'ONE' AND 'TWO'.

Word	Ref prons	Hyp prons	Unilateral pairs	Unilateral V-PA	Bilateral pairs	Bilateral V-PA
abuse	@ b j u z @ b j u s	@ b j u s	@ b j u z → @ b j u s @ b j u s → @ b j u s	90%	@ b j u s → @ b j u z @ b j u s → @ b j u s	90%
ape	@ i p	@ i p A: p @	@ i p → @ i p	100%	@ i p → @ i p @ i p → A: p @	34%
one	w a n	w O n w a n O n e	w a n → w a n	100%	w a n → w a n w a n → w O n w a n → O n e	56%
two	t u: t u	t @	t u: → t @ t u → t @	50%	t u: → t @ t u → t @	50%

```
for each matching word in ref and hyp dictionary do
    # Score pairs by aligning ref and hyp
      pronunciations;
    for each ref variant r do
        for each hyp variant h do
            align r and h;
            acc(r,h) = PDP score of aligned r and h
        end
    end
    # Do greedy search to assign pairs;
    counter = max(number of ref variants, number of
      hyp variants);
    while counter > 0 do
        (selected_h, selected_r, accuracy) = select next
          best from acc;
        pairs(selected_r,selected_h) = accuracy;
        counter–;
    end
    cum_count = 0;
    for each ref var r in pairs do
        for each hyp var associated with r do
            best_phone_acc = pairs(r,h);
            cum_phone_acc = cum_phone_acc +
              best_phone_acc;
            cum_count++;
            if best_phone_acc > single_best_phone_acc
              then
                single_best_phone_acc = best_phone_acc
            end
            if best_phone_acc == 1 then
                cum_word_acc++;
                single_best_word_acc = 1
            end
        end
    end
end
```

Algorithm 1: Bilateral G2P scoring algorithm when comparing pronunciations in the reference ('ref') and hypothesised ('hyp') dictionaries.

IV. EXPERIMENTAL DESIGN

Estimating the G2P accuracy of variant-based dictionaries requires an approach that allows penalising over- or under-generation of pronunciation variants. In order to obtain an empirical result for the task of balancing between G2P accuracy and ASR performance, our experimental setup is designed to mimic previous work carried out in [2]. We used the same setup as discussed in [2]. This section presents a summary of the experimental setup with a pointer to the reference publications.

In previous work, we used two datasets but our experiment make use of one: The South African Directory Enquiries (SADE) corpus [11]. This work focused on four South African languages, namely Afrikaans, English, isiZulu and Sesotho. This corpus provided a word list with their corresponding language of origin to obtain an oracle result (*Ref-LID*). In addition to the *Ref-LID*, we considered three (3) other cases whereby Joint Sequence Models (JSMs) were used for text-based Language Identification (LID) prediction.

G2P models are generated using the full set of the NCHLT-*in-lang* [12] dictionaries as training data. Using these models, we automatically generate four hypothesised dictionaries where the most probable pronunciation is produced using language-specific G2P. The only difference among the dictionaries relates to which language is selected when producing the G2P pronunciation. This work experimented with:

- Ref-LID hypothesised dictionary: A G2P-based dictionary obtained from the manually tagged word list of each corpus.
- Single LID hypothesised dictionary: A G2P-based dictionary obtained from single-language tags (where JSM technique classifies words as monolingual only) [13].
- Multi LID hypothesised dictionary: A G2P-based dictionary obtained from multi-language tags (where JSM technique classifies words as multilingual using threshold approach) (see [14] for details).
- All-four hypothesised languages dictionary: A G2P-based dictionary obtained when assuming that each word originates from all four target language sets (Afrikaans, English, Sesotho and isiZulu).

For proper fair comparison, the resulting hypothesised dictionaries were mapped to two reconciled phoneme sets (combined and detailed) due to differences in phonemic

TABLE II

USING STANDARD PHONE ACCURACY APPROACH TO COMPARE BETWEEN UNI- AND BILATERAL SCORING STRATEGY FOR DIFFERENT PRONUNCIATION DICTIONARY APPROACHES ON SADE DATA SET. RESULTS OBTAINED ON 'COMBINED' PHONEME SET.

Dict	S-WA	Unlilateral V-WA	Bilateral V-WA	S-PA	Unilateral V-PA	Bilateral V-PA	Ref._{avg}	Hyp._{avg}
All-four	63.04	59.96	20.49	90.73	89.59	60.03	5.21	3.40
Multi	60.50	57.26	51.44	88.31	86.94	82.20	5.21	1.33
Single	56.30	52.91	53.17	85.11	83.42	83.42	5.21	1.00
Ref-LID	59.40	56.53	55.37	88.06	86.79	85.89	5.21	1.33

TABLE III

COMPARISON BETWEEN 'ALIGNED' AND 'STANDARD' PHONEME ACCURACY OVER UNI- AND BILATERAL SCORING STRATEGY FOR DIFFERENT PRONUNCIATION DICTIONARY APPROACHES ON SADE DATA SET. RESULTS OBTAINED ON 'COMBINED' PHONEME SET.

Dict	S-PA Standard	S-PA Aligned	Unilateral V-PA Standard	Unilateral V-PA Aligned	Bilateral V-PA Standard	Bilateral V-PA Aligned
All-four	90.73	91.39	89.59	90.56	60.03	65.33
Multi	88.31	89.29	86.94	88.20	82.20	83.99
Single	85.11	86.53	83.42	85.11	83.42	85.11
Ref-LID	88.06	89.05	86.79	88.02	85.89	87.14

transcriptions used in the corpora involved. This current work makes use of only the 'combined' phoneme set.

A baseline ASR system was built using the entire SADE corpus [11]. A partition set size of 65% and 35% was used for training and testing, respectively. The baseline system employs a standard Kaldi-based system using a recipe similar to the Babel recipes [15]. We built a context-dependent crossword HMM-based phone recogniser with triphone models and Gaussian mixture models (GMMs). Both speaker-specific transformation and normalisation are performed. 40 features are used after feature reduction through linear discriminant analysis: splicing together seven frames 13-dimensaional MFCCs each. The models produces alignment, which are used to initialise a standard 3-layer deep neural network. Decoding is performed by employing both a flat and *n-gram* trained language model. To avoid out-of-vocabulary token during our ASR training, we include all words not present in the hypothesised dictionaries by extracting their corresponding pronunciations from the original SADE transcribed lexicon. Note that all phonemes are mapped to the 'combined' phoneme set.

A. Performance measures

The G2P accuracy for variant-based dictionaries is analysed using four metrics defined in [1], namely variant-based phone accuracy (V-PA), variant-based word accuracy (V-WA), single-best phone accuracy (S-PA) and single-best word accuracy (S-WA). (Also see [2].)

Given the above G2P performance metrics, phone accuracy is calculated using two approaches: standard phone accuracy and aligned phone accuracy. These two approaches differ as follows: If I represents insertions, C represents correct phone pairs, S represents substitutions, D represents deletions, and N represents the number of phones in the reference word, then

$$\mathrm{N} = C + S + D \tag{1}$$

and using equation 1, we can calculate:

$$\text{standard phone accuracy} = \frac{C - I}{N} \tag{2}$$

$$\text{aligned phone accuracy} = \frac{C}{N + I} \tag{3}$$

Standard phone accuracy (eq. 2) is typically used, but aligned phone accuracy (eq. 3) has the benefit that matching the hypothesis against the reference, or the reference against the hypothesis produces the same result. This is not always the case for standard phone accuracy. ASR accuracy is evaluated in terms of the word error rate (WER) metric by aligning a recognised word string against the correct word string and computing the number of substitutions (S), deletions (D), insertions (I) and the number of words in the correct sentence (N).

V. ANALYSIS AND RESULTS

A. Unilateral versus Bilateral analysis

Unilateral (one-sided) G2P analysis evaluates pronunciation variants in the hypothesised lexicon against all the variants in the reference dictionary to estimate the best matching variants. This method does not penalise techniques that over-generate pronunciation variants. While more complex to evaluate (the new metrics require an additional alignment per variant pair), the bilateral approach is conceptually a more valid indication of the role that variants play.

When we now re-evaluate phone accuracy (see Table II) the G2P results follow the ASR trends much more closely. In fact, the order of performance – *Ref-LID, Single, Multi* and *All-four* – is a clear match, with the actual results also well correlated, as shown in Fig. 1. We also observe the same trends across dictionary approaches regardless of the language model employed. As expected, a flat LM without lexicon tags performs the worst. In conclusion, bilateral G2P V-PER correlate very closely with ASR WER while unilateral G2P V-PER is a wrong indicator of the ASR performance.

In Table II, we observe that the *All-four* lexicon is affected most by the change in metric: producing a high accuracy with unilateral scoring, this is much lower using bilateral scoring. Similarly, the G2P accuracies of the *Multi* and *Ref-LID* dictionaries also decrease. As expected, the performance of the *Single* dictionary remains the same during unilateral and bilateral scoring.

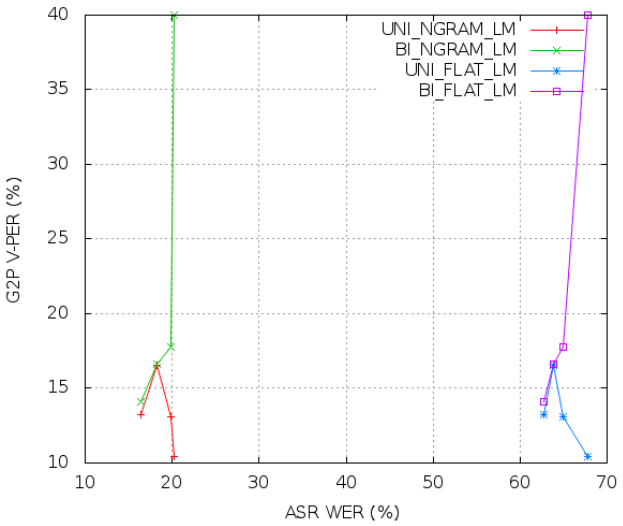

Fig. 1. Comparison between unilateral and bilateral V-PA against ASR WER of the flat and *n-gram* trained LM without variant-tagged lexicon. UNI_NGRAM_LM and BI_NGRAM_LM represent unilateral and bilateral V-PA analysis for decoding performed using *ngram* trained LM. UNI_FLAT_LM and BI_FLAT_LM represent unilateral and bilateral V-PA analysis for decoding performed using flat LM.

B. Additional parameters

In this section, we provide other parameter choices that can affect the performance of the measure. These parameters include:

- Scoring matrix (weighted vs uniform): This matrix is used during dynamic programming-based variant matching to weigh individual phone errors based on how easily two phones are confused. (For example, confusing /i/ with /@/ typically less serious than confusing /i/ with /k/, a rarer occurance.)
- Word frequency: This represents the expected token frequency per word, and can be used to weigh word-based errors with frequency information, for analysis purposes. This measure requires an expected word frequency distribution, typically estimated from the training data.
- Reference length (aligned vs standard): In Section IV-A, we discussed two approaches (standard and aligned) to estimate phone accuracy. An aligned phone accuracy provides the benefit that matching the reference against the hypothesis or vice versa gives the same result. To observe its effect, we analyse the accuracy of the aligned phone accuracy approach with results obtained in Section V-A. Table III shows the G2P accuracy obtained using aligned phoneme accuracy, calculating use uni- and bilateral scoring, respectively. Results show an approximate 1% difference, but a similar trend to the one discussed in Section V-A above. This distinction is therefore not very important for the task studied.

VI. Conclusion

Many G2P techniques (such as JSMs) allow a varying number of variants to be generated. To optimize ASR results, systems must be trained and tested using the same dictionary, which is time-consuming and computationally expensive, while measuring G2P accuracy is computationally inexpensive.

This paper focused on how variants are dealt with during accuracy calculation, which in turn has a significant effect on measured G2P performance. Based on the existing G2P accuracy measures, where variants in the hypothesised dictionary are not penalised sufficiently, we propose a straightforward technique to measure G2P accuracy when both the reference and hypothesised dictionaries contain variants. This technique automatically penalises a dictionary for over- or under-generating variants. Using this new metric, if we order the dictionaries according to G2P performance (best to worst), we obtain an ordering that correlates with the actual ASR performance observed.

Finally, this technique can be used to set variant thresholds for unseen words, based on accuracies observed on a small seen subset of the pronunciation dictionary.

VII. Acknowledgements

We would like to acknowledge Charl van Heerden for his assistance with the ASR experiments, as well as Ulrike Janke for her editing assistance.

This work was partially supported by the National Research Foundation (NRF). Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and therefore the NRF does not accept any liability in regard thereto.

References

[1] M. H. Davel, C. J. van Heerden, and E. Barnard, "G2P variant prediction techniques for ASR and STD," in *Proc. Fourteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013, pp. 1831–1835.

[2] O. Giwa and M. H. Davel, "The effect of language identification accuracy on speech recognition accuracy of proper names," in *Proc. Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2017, accepted for publication.

[3] M. Lehr, K. Gorman, and I. Shafran, "Discriminative pronunciation modeling for dialectal speech recognition," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 1458–1462.

[4] R. G. Brunet and H. A. Murthy, "Impact of pronunciation variation in speech recognition," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, 2012, pp. 1–5.

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict, [Online; accessed 27-September-2017].

[6] W. Fisher, "NIST Syllabification software," [Available at: ftp://jaguar.ncsl.nist.gov/pub/].

[7] S. Hahn, P. Vozila, and M. Bisani, "Comparison of grapheme-to-phoneme methods on large pronunciation dictionaries and lvcsr tasks," in *Proc. Thirteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012, pp. 2538–2541.

[8] D. Jouvet, D. Fohr, and I. Illina, "Evaluating grapheme-to-phoneme converters in automatic speech recognition context," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4821–4824.

[9] B. Hixon, E. Schneider, and S. L. Epstein, "Phonemic similarity metrics to compare pronunciation methods," in *Proc. Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[10] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proc. Third International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTUÍ2)*, 2012, pp. 68–75.

[11] Thirion, Jan W.F. and van Heerden, Charl and Giwa, Oluwapelumi and Davel, Marelie H., "The South African Directory Enquiries (SADE) corpus," in preparation.

[12] M. H. Davel, W. D. Basson, C. van Heerden, and E. Barnard, "NCHLT Dictionaries: Project Report," Multilingual Speech Technologies, North-West University, Tech. Rep., May 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/home

[13] O. Giwa and M. H. Davel, "Language identification of individual words with joint sequence models," in *Proc. Fifteenth Annual Conference of the International Speech Communication Association, INTERSPEECH*, 14-18 September, Singapore, 2014, pp. 1400–1404.

[14] O. Giwa and M. Davel, "Text-based language identification of multilingual names," in *Proc. Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015, pp. 166–171.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, IEEE Catalog No.: CFP11SRW-USB.