Unique n-Phone Ranking Based Spoken Language Identification Using Phone Lattices

Amalia Zahra and Julie Carson-Berndsen
CNGL, School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
Email: amalia.zahra@ucdconnect.ie, julie.berndsen@ucd.ie

***Abstract*—This paper presents a novel approach to language identification (LID) which is unique *n*-phone ranking (UnPR) calculated from *p*-best phone recogniser outputs. The idea underlying this approach is to create a list consisting of a set of unique *n*-phones considered to be the approximate identity of a certain language. Thus, every language in the target set has its own list. These lists support the process of distinguishing one language from another. The novel part of the proposed approach lies in the procedures for generating such lists and utilising them to perform LID tasks. Compared to our previous work, the added value of the work presented in this paper is taking phone lattices into account to investigate whether the updated UnPR can improve the LID performance. Besides the UnPR from our previous work, parallel phone recognition followed by language modelling (PPRLM) is also used as a baseline system. Furthermore, a fusion combining the UnPR system with a number of configurations is carried out. The experiments show that the proposed approach complemented with lattices improves LID accuracies. The amount of time spent is also significantly faster, which is ~9.5 times faster than the PPRLM system.**

***Index Terms*—spoken language identification, phonotactic approach, phone lattices**

I. Introduction

Spoken language identification (LID) is the task of determining the language of an unknown speech utterance. Approaches applied are typically classified into two categories depending on the types of features used for identification: acoustic or phonetic. Acoustic-based LID relies solely on acoustic features of the speech utterances without any linguistic resources involved, such as transcriptions, phonetic annotations, lexicons, etc. Approaches that utilise such linguistic resources are categorised as phonetic-based LID. Acoustic-based approaches have advantages over phonetic-based ones due to the fact that less resources are required to be able to develop an LID system, especially for under-resourced languages. On the other hand, phonetic-based approaches also have advantages namely that higher level information (i.e. linguistic resources) is taken into account thus a reasonable linguistic analysis can be derived to justify the system performance.

Manuscript received August 11, 2013; revised November 19, 2013.

Even though phonetic-based LID requires a number of linguistic resources, the fact is that most languages do not have such resources. Therefore, a form of phonetic estimation is necessary to capture some sort of linguistic knowledge from the data of those languages. This is the underlying concept of phonetic-based approaches that have been widely implemented for LID tasks. The main idea of phonetic-based approaches is to utilise at least one phone recogniser, which is built for the language that has sufficient linguistic resources, to estimate the phone sequences of the data of all the target languages covered by the LID system. These phone sequences are processed further to develop an LID system, which is also known as phonotactic LID since phone sequences are mostly utilised.

In automatic speech recognition (ASR), an unknown speech utterance is typically decoded into a single sequence (i.e. phone sequence). A number of studies have used this output to perform LID. The work in [1] used phone recogniser outputs to build a hierarchical multilayer perceptron for five-language identification tasks; while the work in [2] utilised 1-best outputs from multiple phone recognisers to develop an approach that takes into account time alignment information by considering time synchronous cross-decoder phone co-occurrences. A recogniser can generate not only 1-best output, but also *p* possible phone sequences (*p*-best list) for a single speech utterance in the representation of a lattice. Thus, richer information regarding the utterance is taken into account. Lattices are not only utilised in ASR [3], but also in LID. The work in [4] used lattices as resources to obtain the counts of tokens (i.e. phones) and *n*-grams which were subsequently used as features to train a support vector machine (SVM) based LID system complemented with a channel compensation technique named nuisance attribute projection (NAP) [5]. Lattices were also used in [6] to improve LID performance by removing pronunciation and grammar constraints to provide a richer set of decoding alternatives and context for embedded LID in large vocabulary continuous speech recognition (LVCSR) system. An LID improvement was also achieved in [7] by using lattices to obtain better maximum likelihood estimates in the language models employed in their LID system.

The work presented in this paper is a continuation of [8], whose novelty lies in the way the phone sequences of speech utterances are processed, which will be explained

doi: 10.12720/ijsps.1.2.190-195

briefly in this paper. Its key advantage is that it utilises multiple lengths of *n*-phones to capture a broader context of an utterance without having to worry about high dimensionality issue, as experienced in [9]. Moreover, it runs faster than some of state-of-the-art phonetic-based approaches [8], [10].

Compared to [8], the new value added into our system is the usage of phone lattices. The idea is to enrich the unique phone lists used in the UnPR and investigate whether lattices can further improve the original UnPR system accuracy. Furthermore, a performance comparison is also carried out between the lattice-based UnPR and the well-known LID approach called parallel phone recognition followed by language modelling (PPRLM) [10], [11] as a baseline.

The remainder of this paper is structured as follows. Section II briefly describes the baseline system (PPRLM). Section III explains the proposed approach (UnPR) covering the original concept of UnPR and the scenario where lattices are taken into account. As a number of configurations are applied in this paper, a fusion method over those configurations is carried out on the UnPR system, which will be described in Section IV. Section V details all of the data and the configurations used in the experimental evaluation. Section VI presents and discusses the results. Finally, conclusions are drawn and future work is outlined in Section VII.

II. Parallel Phone Recognition Followed by Language Modelling (PPRLM)

As a baseline system, parallel phone recognition followed by language modelling (PPRLM) [10] is used. The idea of PPRLM is to utilise a number of front-end phone recognisers to decode the speech data of each target language. The reason for this is that the sounds of the target languages do not always occur in the one language used to train the phone recogniser. Therefore, multiple phone recognisers, each trained on different language, are expected to cover more sounds in the target languages (if not all). The outputs of each recogniser are then used to build language models (i.e. interpolated *n*-gram LMs), one corresponds to each target language. Thus, the training process produces $M \times N$ language models where M corresponds to the number of target languages and N the number of phone recognisers.

During evaluation, an unknown speech utterance is firstly decoded using each of the N phone recognisers. The phone sequence obtained is then tested with each of the M language models, which produces $N \times M$ log likelihoods. The log likelihoods are averaged over the N recognisers for each target language. This results in M overall language log likelihoods. The language that produces the highest overall log likelihood is decided as the hypothesised language of the unknown speech utterance.

Since lattices are used in the proposed approach, which will be explained in Section III, the baseline system also uses lattices to achieve fairness. It means that both training and test data consist of *p*-best phone sequences for every speech utterance.

III. Unique N-Phone Ranking (UnPR) Using Lattices

As mentioned in Section I, the work presented in this paper is an extension of the unique *n*-phone ranking (UnPR) [8]. Thus, the original concept of the UnPR will be explained in less detail in this paper and more focus placed on the usage of lattices in the effort to investigate if they can improve the accuracy of the LID system. Therefore, the previous system is also used as a baseline (i.e. UnPR without lattices) in addition to the lattice-based PPRLM system.

UnPR deals with a novel way to process the phone sequences generated by a front-end phone recogniser. While in PPRLM the phone sequences are used to create a set of LMs, in UnPR they are used to create a unique list of the most frequent *n*-phones (ranked based on the number of occurrences) with certain lengths for each target language (henceforth called *unique ranked list*). All *n*-phones in the list of a language are guaranteed to be not overlapping with those in the lists of the remaining languages. Thus, this unique ranked list is considered to be the discriminative list that distinguishes a language from the others. This is part of the training process in UnPR where its finer explanation will be presented in Subsection III.A and followed by the evaluation process in Subsection III.B.

A. Training

Four steps are performed in the training process, given M sets of phone sequences generated by a phone recogniser (M target languages). Steps 1, 2, and 4 are carried out on the data of each target language independently while step 3 requires the data of all the target languages.

1. Given a number of different lengths r, *n*-phones are extracted from each phone sequence and then sorted by their number of occurrences in descending order, which means the most frequent *n*-phone occupies the first rank.
2. The top k% of the sorted *n*-phones are extracted. This set represents the most frequent unique *n*-phones of a language. k is assigned with different values in this paper for the purpose of LID performance comparison.
3. To get a unique ranked list for each target language, the list obtained from the previous step is compared with that of all other target languages. The overlapping *n*-phones between the language and each of all other target languages are removed from the list.
4. Finally, for every language, each *n*-phone is assigned with a value that represents its rank-based weight. The weight is decreasing from the top to the bottom rank. This means that the most frequent *n*-phone is assigned with the largest weight, which indicates a high significance.

By the end of the training process, M unique ranked lists (i.e. list of *n*-phones and their weights) are ready to be used for LID evaluation.

In our previous work [8], only the most likely sequence (i.e. 1-best output) generated by a phone recogniser was taken into account. However, a phone recogniser is capable of generating a lattice from which a number of possible sequences for a single speech utterance (henceforth called *p*-best list) is extracted.

As the focus of this paper is to investigate if the usage of lattices can further improve LID accuracy, all the above four steps are carried out on the *p*-best recognition outputs. This means that more data is taken into account which results in more accurate unique ranked lists. The reason underlying this assumption is that the more the data brought into the training process, the more accurate the selection applied to generate the unique *n*-phone lists.

B. Evaluation

During evaluation, two steps are carried out, given a phone sequence of an unknown speech utterance generated by a phone recogniser.

1. For every *n*-phone in the utterance that exists in the unique ranked list of a particular target language, extract its rank-based weight and then sum them up in a weighted sum scoring procedure to obtain the score for that language. As mentioned in step 1 of the training process, different lengths *r* are used (e.g. *r*=3 represents 3 lengths of phone sequences: bi-, tri-, and pentaphones). Thus, there are *r* additional weights assigned for the *r* sets of *n*-phones (henceforth called *length-based weights* to distinguish them from the rank-based weights). Typically, the longer the *n*-phone, the larger the length-based weight should be applied to the set, as longer *n*-phones are more discriminative. This is accommodated in the weighted sum scoring formula described in [8].
2. Since the weighted sum score is computed for each target language, then there are *M* scores obtained where each score corresponds to a target language. Finally, the hypothesised language of the unknown speech utterance is then decided as the one whose list produces the maximum score.

Lattices are used not only in the training process, but also in the evaluation. The rank-based weights mentioned in step 1 are counted over the *p*-best recognition outputs of the unknown speech utterance. Such outputs seem as if the unknown utterance had longer duration as more sequences are taken into account. This might be beneficial for LID task as the longer the speech, typically the better the LID accuracy.

IV. Fusion

In this paper, three phone recognisers are used which denote three different configurations. UnPR is carried out using each of these configurations where the results are fused. The fusion takes standard deviation into account to investigate the level of significance of the winning score (step 2 in Subsection III.B) among the scores of all other target languages.

The level of significance (L) of a winning score (S) produced with a certain configuration (c) is defined in

$$L(S) = \frac{S - \sigma_c}{\mu_c} \quad (1)$$

where σ_c and μ_c denote the standard deviation and the average of M scores (M represents the number of target languages), respectively, given a certain configuration. The final hypothesised language is the one that has the highest level of significance.

V. Experimental Evaluation

Three types of evaluation are carried out: PPRLM using lattices, UnPR using 1-best recognition output (both are baselines), and UnPR with lattices (proposed). The data used in the experiments is 2007 NIST Language Recognition Evaluation Test Set [12] which consists of conversational telephone speech segments in 22 languages: Arabic, Bengali, Chinese Cantonese, Chinese Mandarin, Chinese Min, Chinese Wu, English, Farsi, French, German, Hindustani, Indonesian, Italian, Japanese, Korean, Punjabi, Russian, Spanish, Tagalog, Tamil, Thai, and Vietnamese. However, only 21 languages are used in the experiments (Punjabi was excluded due to the small amount of data which creates imbalance in training or testing). There are three nominal durations in this corpus: 3s, 10s, and 30s.

The remaining data is split into 80% for training and 20% for testing. The length of the training data for each target language is mostly ~46m with the exception of Chinese Mandarin, Russian, Tamil, and Vietnamese where the training data is ~1.5h each; and the length of the English, Hindustani, and Spanish training data is ~2.3h each.

The test set has similar distribution, which means the larger the training set for a certain target language, the larger the test set for that language (due to the 80%-20% split). Hence, most of the speech data in the test set are 15 or 16 utterances for each nominal duration, which results in 45 or 48 utterances for each target language. However, Chinese Mandarin has 90 utterances (30 for each duration); Russian, Tamil, and Vietnamese have 93 utterances each (31 for each duration); English, Hindustani, and Spanish have 138 utterances each (46 for each duration).

There are three single-language phone recognisers used in the experiments. They are Czech, English, and Russian phone recogniser, developed by Brno University [13]. They implement neural network classifiers and Viterbi algorithm to decode the phone sequence. The toolkit to build language models (for PPRLM) and generate *p*-best list from lattices (for both PPRLM and UnPR) is SRI Language Modelling (SRILM) Toolkit [14].

The language model used in the PPRLM is interpolated 5-gram distribution where the weights applied to unigram, bigram, trigram, tetragram, and pentagram are 0.05, 0.1, 0.15, 0.3, and 0.4, respectively. On the other hand, the UnPR uses biphones, triphones, and pentaphones simultaneously (i.e. *r*=3). Thus, the length-based weights applied to the biphones, triphones,

and pentaphones are 0.2, 0.3, and 0.5, respectively (step 1 in Subsection III.B). Moreover, three different percentages are used to extract the top *n*-phones from the ranked list (the value *k* in step 2 in Subsection III.A): 50%, 75%, and 100%.

Since lattices are used in the experiments, different numbers of top phone sequences (*p*-best list) given a speech utterance are taken into account for the purpose of LID accuracy comparison. Thus, 10-, 20-, 30-, 40-, and 50-best list are included in the UnPR system; while the PPRLM system only uses 50-best list. Recall that the UnPR system without using lattices is also developed as a baseline, in addition to the PPRLM system.

VI. Results and Discussion

Four types of experiments are carried out in this paper. Firstly, the effect of using lattices in the UnPR is investigated. The outcomes are also compared to the PPRLM system. A fusion of different configurations of UnPR is analysed. Finally, an extra evaluation is carried out using both UnPR and PPRLM (with 50-best list) to investigate their capability in identifying linguistically close languages. Therefore, the results of these four experiments are consecutively presented in the four following subsections.

A. 1-Best and Lattice Based UnPR

As mentioned in Section V, three phone recognisers are used: Czech, English, and Russian phone recogniser. The investigation is carried out on the results of the UnPR using 1-best recognition output (without lattices) and that using lattices with five *p*-best lists (different values for *p*), which are 10-, 20-, 30-, 40-, and 50-best list.

Nine charts are presented in Fig. 1, arranged based on their nominal durations (i.e. 3s, 10s, and 30s) as rows and the amount of top *n*-phones extracted for the unique ranked lists (i.e. top-50%, -75%, and 100%) as columns.

As illustrated in each of the charts, the LID accuracies tend to improve along with the larger *p*-best list used. One possible reason for this positive tendency is the fact that more data is included in the process of creating the unique ranked lists even though the number of speech data for training is considered small compared to other LID studies that have been carried out [1], [2], [4]. The best LID accuracies achieved in this experiment are those obtained using all unique *n*-phones (100%) in the unique ranked lists, identified during training. They are 72.4%;92.0%;86.7% (3s;10s;30s), 97.0%;96.8%;95.6%, and 71.3%;91.8%;85.9% using Czech, English, and Russian phone recogniser, respectively (highlighted in circle in Fig. 1). All these percentages (except for the 30s speech evaluation using the English recogniser) are achieved when lattices are taken into account. Hence, the usage of lattices does improve the overall LID performance of the UnPR system.

B. Lattice Based UnPR and PPRLM

Since the best accuracies achieved in the previous experiments are those produced by the English phone recogniser, they are compared with the results of PPRLM using interpolated 5-gram LMs on 50-best lists generated by Czech, English, and Russian phone recogniser. Table I presents the LID accuracies of both approaches for 3s, 10s, and 30s speech.

TABLE I. LID Accuracies of UnPR (English Recogniser) and PPRLM (50-best; interpolated 5-gram LMs)

approach	3s	10s	30s
UnPR	97.0%	96.8%	95.6%
PPRLM	97.9%	96.2%	94.3%

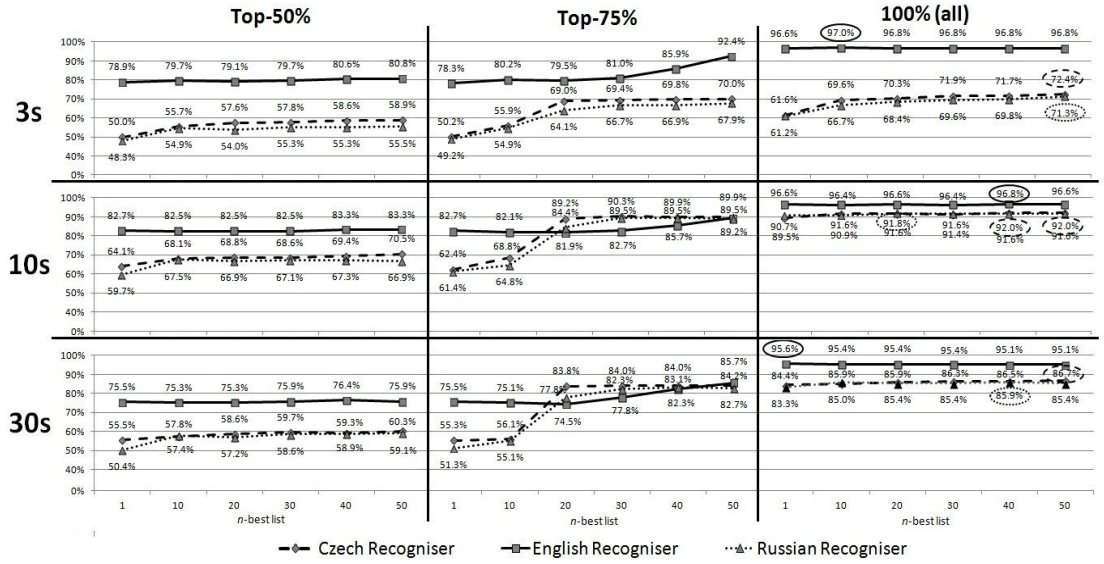

Figure 1. LID accuracies of unique *n*-phone ranking (UnPR) system.

Compared to PPRLM, the LID accuracies achieved by UnPR are slightly higher, except for the evaluation of the 3s speech segments. Even though the improvement gained is not significant, the UnPR system runs

significantly faster than PPRLM. On the test set, the UnPR system is ~9.5 times faster than PPRLM system for 50-best configuration (Intel(R) Core(TM) i5-2410 M CPU 2.3 GHz processor). The reason for this is that UnPR deals with a simple procedure of determining the language of an unknown speech based on a set of unique ranked lists created in the training process. It is a simple yet logical and efficient approach for LID tasks.

C. Fusion with UnPR Systems

Since there are three phone recognisers used in the experiments, a fusion is carried out on their outputs. The final decision on the hypothesised language of an unknown speech utterance is performed based on the level of significance of each winning score (from each phone recogniser), as explained in Section IV. Table II shows the LID accuracies of the fusion system using the five *p*-best lists and all (100%) the unique *n*-phones for each target language.

Unlike the results presented in Fig. I which indicate positive correlation between the number of possible phone sequences used (i.e. *p*-best list) and the LID accuracy, the overall results shown in Table II do not change significantly when the *p*-best list is enlarged. The reason underlying this outcome might be due to the fact that the performance of the UnPR system using only English phone recogniser is already high (see Fig. I) thus its fusion with the results of the two other recognisers does not result in a significant impact even though the *p*-best list is enlarged. However, the fusion slightly improves the identification of the 10s and 30s speech segments from 96.8% (without fusion, see Subsection VI.A) to 97.0% and from 95.6% (without fusion) to 96.6%, respectively.

In the case where the phone recogniser used does not have such high accuracies on its own (e.g. built with limited data), it is likely that fusion will have a greater impact and this approach deserves further investigation. Hence, the outputs of the English phone recogniser are excluded. The results of a fusion using Czech and Russian phone recognisers are presented in Table III.

As shown in Table III, the LID accuracies improve for all the three nominal durations, compared to those produced by Czech and Russian phone recogniser independently (Subsection VI.A). The best accuracies achieved by fusing the LID outputs of those two phone recognisers are 77.4%, 93.5%, and 96.0% for 3s, 10s, and 30s speech, respectively.

TABLE II. LID ACCURACIES OF THE FUSION SYSTEM USING CZECH, ENGLISH, AND RUSSIAN PHONE RECOGNISER

p-best list	3s	10s	30s
10-best	96.2%	**97.0%**	95.8%
20-best	96.6%	**97.0%**	96.0%
30-best	96.8%	96.4%	95.8%
40-best	96.6%	96.2%	**96.6%**
50-best	96.6%	96.2%	96.2%

TABLE III. LID ACCURACIES OF THE FUSION SYSTEM USING CZECH AND RUSSIAN PHONE RECOGNISER

p-best list	3s	10s	30s
10-best	75.5%	92.6%	88.0%
20-best	77.0%	93.2%	**96.0%**
30-best	76.6%	**93.5%**	89.2%
40-best	76.4%	**93.5%**	90.1%
50-best	**77.4%**	93.0%	89.9%

D. Linguistically Close Languages (UnPR and PPRLM)

All the previous three experiments are carried out for all 21 target languages. Among these languages, there is a number of linguistically close language sets according to the language family to which each one of them belongs. It might be interesting to investigate the capability of UnPR and PPRLM in terms of identifying languages within their own language family.

Of the 21 target languages included in this paper, six sets of linguistically close languages are defined. Table IV shows the six sets and their LID accuracies within their own sets for 3s, 10s, and 30s speech. Note that the percentages in brackets represent the accuracies of the PPRLM system, while those not in brackets represent UnPR.

From the results shown in Table IV, it is proven that both UnPR (proposed) and PPRLM (baseline) achieve high accuracies in identifying languages within their own family with slight differences in a few cases (shaded cells in Table IV). Apart from the fact that both approaches are capable of distinguishing such languages, this promising performance might also be due to the few number of languages involved in each set (i.e. only 2 to 4 languages per set).

TABLE IV. LID ACCURACIES FOR LINGUISTICALLY CLOSE LANGUAGES WITH UNPR AND PPRLM

language family	languages	3s	10s	30s
Germanic	1. English 2. German	100% (100%)	100% (100%)	98.4% (98.4%)
Romance	1. French 2. Italian 3. Spanish	100% (100%)	98.7% (98.7%)	98.7% (100%)
Indic	1. Bengali 2. Hindustani	96.7% (96.7%)	98.4% (98.4%)	98.4% (98.4%)
Altaic	1. Japanese 2. Korean	100% (100%)	100% (100%)	100% (100%)
Sino-Tibetan	1. Chinese Cantonese 2. Chinese Mandarin 3. Chinese Min 4. Chinese Wu	98.7% (98.7%)	97.4% (98.7%)	97.4% (96.1%)
Austronesian	1. Indonesian 2. Tagalog	100% (100%)	100% (100%)	100% (100%)

percentages: UnPR (PPRLM)

Based on an observation on the results of the four experiments above, there appears to be many cases where LID accuracies degrade for longer speech segments. This is peculiar as typically the longer the speech, the more accurate the LID system. A few random checks were done and it was found that some utterances with longer

duration have longer non-speech segments (e.g. filler, laughter, etc) compared to shorter utterances. This might be the reason such degradation happens. However, it is still a hypothesis, hence further investigation is required.

VII. Conclusions and Future Work

This paper has shown that using lattices in the proposed approach (UnPR) has a tendency to improve the LID accuracy. By taking lattices into account, the unique ranked list, which is the most important part in the approach, can be created more accurately as more data is included. The more the data included in the process, typically the greater the amount of phone sequence overlaps between languages (which is discarded from the list), which makes the creating of the unique ranked list more selective.

Moreover, UnPR achieves a slight improvement compared to PPRLM in most of the nominal durations. However, the amount of time spent to run UnPR is significantly faster than that spent to run PPRLM. The difference is ~9.5 times faster on the evaluation of the test set. It can also be concluded that a fusion on the UnPR systems has a positive impact when the accuracies achieved in the individual systems are not so accurate. However, if one of the results that will be fused is already highly accurate prior to fusion, it is expected that the accuracies might not necessarily improve further. Finally, both UnPR and PPRLM are able to identify linguistically close languages within their own language family with slight differences in their accuracies.

As there appears to be cases where the LID accuracies degrade for longer speech segments, it is necessary to carry out further investigation about the reason underlying this phenomenon. One way that might be useful to address this issue is to apply a speech activity detection (SAD) procedure in the system so that only the segments that are considered as actual speech are taken into account (and discarding non-speech segments that do not contain any linguistic information). Moreover, our future work will also involve integrating the UnPR approach into our experimental framework for LID.

Acknowledgement

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin (UCD). The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

References

[1] D. Imseng, M. M. Doss, and H. Bourlard, "Hierarchical multilayer perceptron based language identification," in *Proc. 11th Annual Conference of the International Speech Communication Association*, 2010.

[2] M. Penagarikano, A. Varona, L. J. Rodriguez-Fuentes, and G. Bordel, "Improved modeling of cross-decoder phone cooccurrences in SVM-based phonotactic language recognition," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2348-2363, 2011.

[3] J. Duchateau and K. Demuynck, "Evaluation of phone lattice based speech decoding," in *Proc. Interspeech*, 2009, pp. 1219-1222.

[4] F. S. Richardson and W. M. Campbel, "NAP for high level language identification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4392-4395.

[5] W. M. Campbel, D. E. Sturim, P. Torres-Carrasquilo, and D. A. Reynolds, "A comparison of subspace feature-domain methods for language recognition," in *Proc. Interspeech*, 2008, pp. 309-312.

[6] Y. Shan, Y. Deng, J. Liu, and M. T. Johnson, "Phone lattice reconstruction for embedded language recognition in LVCSR," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 1, 2012.

[7] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proc. Interspeech*, 2004.

[8] A. Zahra and J. Carson-Berndsen, "Unique n-phone ranking based spoken language identification," in *Proc. IEEE 5th International Conference on Computational Intelligence, Communication Systems and Networks*, 2013, pp. 219-224.

[9] H. Li, B. Ma, and C. H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271-284, 2007.

[10] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 31-44, 1996.

[11] L. Wang, E. Ambikairajah, and E. H. C. Choi, "Multi-lingual phoneme recognition and language identification using phonotactic information," in *Proc. IEEE 18th International Conference on Pattern Recognition*, vol. 4, 2006, pp. 245-248.

[12] A. F. Martin and A. N. Le, "NIST 2007 language recognition evaluation," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[13] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 325-328.

[14] A. Stolcke, "SRILM an extensible language modeling toolkit," in *Proc. International Conference on Spoken Language Processing*, vol. 2, 2002, pp. 901-904.

Amalia Zahra was born in Jakarta, 27th June 1986. She undertook her undergraduate degree at the School of Computer Science, University of Indonesia (2004 – 2008). Her undergraduate thesis was about the development of an Indonesian pronunciation dictionary for a continuous speech recognition system. Afterwards she became a research assistant for a year in the same university where she and her colleague developed an Indonesian large vocabulary continuous speech recognition system. The system was exhibited at the Indonesian Creative Products Expo in 2009. Later that year onwards, she has been pursuing her PhD at the School of Computer Science and Informatics, University College Dublin, Ireland under the supervision of Prof. Julie Carson-Berndsen (co-author).Her research interests have been mainly in the area of language technology, especially the ones related to speech. They include automatic speech recognition, speech synthesis, spoken language identification, phonotactics, signal processing, and machine learning. Furthermore, she is also interested in other fields, still in the area of language technology, such as machine translation and information retrieval.

Julie Carson-Berndsen was born in Dublin in 1963. She completed her undergraduate degree at Trinity College Dublin (BA Hons (Mod)) in German and Mathematics in 1986. She obtained her doctorate from University of Bielefeld in Computational Phonology in 1993. Currently she is a Professor at University College Dublin. Her research interests cover speech technology and computational linguistics.