Processing Spoken Lectures in Resource-Scarce Environments

Charl J. van Heerden
Human Language Technology Competency Area
CSIR Meraka Institute
Pretoria 0001, South Africa
Email: cvheerden@gmail.com

Pieter de Villiers, Etienne Barnard, Marelie H. Davel
Multilingual Speech Technologies
North-West University
Vanderbijlpark 1900, South Africa
Email: {pieterdevill, marelie.davel, etienne.barnard}@gmail.com

***Abstract*—Initial work towards processing Afrikaans spoken lectures in a resource-scarce environment is presented. Two approaches to acoustic modeling for eventual alignment are compared: (a) using a well-trained target-language acoustic model and (b) using an acoustic model from another language, in this case American English. We show that while target-language acoustic models are preferable, similar performance can be achieved by repeatedly bootstrapping with the American English model, segmenting and then adapting or training new models using the segmented spoken lectures. The eventual systems perform quite well, aligning more than 90% of a selected set of target words successfully.**

I. Introduction

The use of speech recognition in lecture rooms has long been understood as a potentially rewarding endeavour [1]. Such systems could be used to generate real-time captions, or multimedia transcripts of lectures, and would clearly be of major value to students with hearing disabilities, or students who are not fluent in the language of the lecture. An awareness of these benefits has prompted significant research and development efforts, and research groups in the USA, Italy, Japan, Canada and the United Kingdom have all demonstrated promising systems (for an updated list of relevant publications, see http://liberatedlearning.com).

In the developing world, the benefits of systems for lecture transcription are potentially even greater, given the lower literacy levels and greater multilingualism that are generally prevalent. However, resource constraints have to date limited lecture-transcription systems to the languages of the developed world, and it is clear that developing such systems with the limited resources of the developing world will require substantial innovations. In the current paper, we propose a few steps in that direction.

In particular, we investigate ways to employ speech-processing tools in order to utilize such resources as may reasonably be expected to exist in less-resourced environments, such as audio recordings of variable quality and (potentially inaccurate) transcriptions of some of those recordings. Our focus is on two processing steps that are basic building blocks in the processing of spoken lectures, namely *alignment* and *term detection*. These processes support further stages of lecture transcription (such as speaker adaptation) and are also useful for purposes such as the production of multimedia transcripts or the development of tools for searching through recordings.

The main contributions of this work are

- the definition and evaluation of a process for the alignment of recordings and transcriptions, which is applicable in under-resourced environments, and
- a report on the first results achieved on a new Afrikaans corpus of recorded university lectures.

In Section II below, we provide background on several topics related to our research, and Section III describes the approach we followed for the alignment and indexing tasks. Our results are reported in Section IV, and in Section V we summarize our findings and provide context on the additional work that is required to develop practically useful lecture-transcription systems in under-resourced languages.

II. Background

Speech processing for under-resourced languages has been receiving increasing attention in recent years, as awareness of the potential role of speech technology in addressing developmental challenges has grown [2]–[4]. The approaches that have been followed can be arranged on a continuum from systems that lean heavily on systems developed for well-resourced languages ([5], for example, uses an English speech-recognition system to recognize speech in ten different languages through phone mappings) to those that assemble all the resources to develop independent systems in the targeted under-resourced language. Van Heerden *et al.* [6] demonstrated significant benefits in accuracy achievable with the latter class of approaches for small-vocabulary speech recognition – but, of course, at the cost of requiring all the necessary resources.

As mentioned in Section I, most work on lecture transcription to date has been quite resource intensive. For example, the system for American English described in [7] used an acoustic model that had been trained on 121 hours of speech and language models trained with more than 6 million words of speech; more than 200 hours of transcribed lectures were then used for the refinement and evaluation of that system. Impressive results have been achieved in such efforts: the MIT team reported word error rates as low as 17% for a complete recognition task in [7], and have demonstrated that very useful applications can be developed around a recognizer

with that level of accuracy. Similarly, a note-taking system for Japanese lectures – described in [8] – was able to improve the producitivity of human editors by a large margin. That system was based on acoustic models trained with the 658-hour Corpus of Spoken Japanese.

A crucial aspect of our approach is its reliance on whatever transcriptions are available; such "approximate" transcriptions have received significant attention since the pioneering work of Lamel *et al.* [9]. A three step approach – consisting of (1) data segmentation, (2) word-based alignment and (3) filtering – has gained wide acceptance, and several algorithms have been proposed for each of these steps: see [10] for a recent review, which specifically addresses the challenges of under-resourced languages. The filtered data is subsequently used to train or adapt acoustic models, using standard approaches for supervised training.

Recently, the use of unsupervised training has received significant attention [11]–[13]. Acoustic models are constructed from untranscribed data, and these are improved in an iterative fashion. Two types of approaches are used: (1) using no resources apart from the acoustic data, and (2) when a language model is available to guide recognition. In the first case only preliminary results have been achieved while requiring very large corpora as input [12], [13]. In the latter case, working systems are achieved without the use of transcribed audio, but then the existence of good language models is an important prerequisite for system development [11].

III. Approach

Our Afrikaans spoken lectures corpus currently consists of 12 recorded lectures (9 male and 3 female) in various domains. (We are planning on extending the corpus significantly in the coming months.) These lectures vary in duration from less than 5 minutes to approximately 45 minutes. A variety of acoustic conditions prevail in the lectures – generally, static microphones were employed in somewhat noisy classrooms. While all lectures have corresponding transcriptions that vary from approximate to accurate, no segmentation information is available (that is, no information is supplied with regard to which portions are transcribed and where they start or stop).

A. Text normalization

The lectures in question were transcribed by different transcribers over a period of 4 years, resulting in predictable inconsistencies with regard to transcription protocols for entities such as numbers and abbreviations. Disfluencies, repetitions and filled pauses were also not transcribed consistently (some transcribers would include them in detail, while others would transcribe as if the speech was fluent and grammatically correct). Spelling mistakes were also common, which can be detrimental to automatic pronunciation prediction approaches. Another problem – expected to be common in many resource scarce environments – is the frequent use of English words and informal speech during lectures.

All of the abovementioned problems need to be addressed in this domain, not only for the purpose of alignment using acoustic models, but also to enable accurate and efficient indexing of such a corpus.

The specific approach we followed to perform text normalization entailed a six-step process:

- The entire corpus was spell-checked using an Afrikaans spellchecker.
- Proper names were identified by inspecting all capitalized words. Once identified, pronunciations were created manually.
- Abbreviations and acronyms were identified by considering all words with 4 or fewer letters, with at most one vowel. Pronunciations for both the spoken as well as the abbreviated form were then created and added to the dictionary.
- Numbers written as digits were normalized to their spoken form where there was no ambiguity. Where ambiguity exists (for example in the pronunciation of "100", where the "one" is often omitted), the number was replaced with a special token, with both corresponding pronunciations being allowed in the dictionary.
- Possible English words were identified in a South African English dictionary. Because there is non-negligible overlap between English and Afrikaans words, both the English pronunciation and an Afrikaans pronunciation (generated by rules if necessary) were retained for such words.
- Pronunciations were automatically generated for any remaining words not in our reference dictionary, using the Default & Refine algorithm [14] and the reference dictionary from [15].

B. Alignment

We wanted to investigate the alignment of the lectures under two conditions: one where we have a well-trained Afrikaans acoustic model available, and another where we try to align the corpus using a well-trained American English acoustic model.

1) NCHLT Afrikaans models: For the first approach, we trained a model using the NCHLT Afrikaans corpus [16], which consists of 185 speakers, with approximately 500 3-5 word utterances per speaker. This amounts to approximately 80 hours of high-bandwidth speech.

The acoustic model was trained on 39 dimensional Mel frequency cepstral coefficients (13 static with cepstral mean normalization, 13 deltas and 13 double deltas). The hidden Markov models (trained with HTK [17]) were standard 3-state left to right tied-state triphone models, with 7 mixtures per state and semi tied transforms. The tied states were created using decision tree clustering. A 14 mixture garbage model was then trained on the entire corpus and combined with this initial model. This model was then used to perform initial alignment, inserting optional garbage markers between words to absorb disfluencies as well as inaccurately transcribed and untranscribed portions.

At this stage, we had initial alignments, with potentially untranscribed or poorly transcribed sections marked by the garbage model. The next step entailed salvaging as much of

the good quality alignments as possible for further retraining. This was accomplished by following an approach described in [10], which is based on a dynamic-programming (DP) phone string alignment procedure. It compares the result of a forced alignment with that of a free decode, using a variable cost matrix, and subsequently identifies accurately transcribed sections of audio and corresponding text. These accurate portions were then in turn used to adapt the NCHLT model on a per-lecture basis, using MAP adaptation, followed by another round of alignment. MAP adaptation was only performed where a lecture was at least 15 minutes in duration (and in those cases we adapted on approximately half of the available speech, retaining the other half for evaluation).

2) WSJ models: In resource-scare environments, large corpora such as the NCHLT Afrikaans corpus are generally not available. While we know that a model trained on data from the target-language is likely to produce better alignments, it is interesting to determine how closely one can approximate the language-specific results when using a well-trained model from a different language.

For this reason, we trained an American English acoustic model using the WSJ corpus and the CMU pronunciation dictionary, using the same parameters as for the NCHLT Afrikaans model. We followed an iterative approach, where the first iteration required remapping of phones from both languages to come up with a common phone set. We employed linguistic knowledge to generate such a mapping. As the transcription conventions used in the CMU dictionary do not model the schwa (/ax/) separately (it is modelled as an unstressed variant of the other vowels that are marked explicitly in the dictionary), we first employed an interpolated phoneme mapping to identify likely occurrences of schwas. Specifically (using ARPABET notation) all the /eh r/, /uh r/, /uw r/, /ih r/, /iy r/ and /er/ samples were mapped to /eh ax r/, /uh ax r/, /uw ax r/, /ih ax r/, /iy ax r/ and /ax r/ respectively and the unstressed /ah/ mapped to /ax/ (retaining stressed /ah/ as /ah/). Once the dictionary was reformatted, each phoneme (or combination of phonemes) was mapped directly to their closest Afrikaans counterparts. 18 of the phonemes could be mapped directly, the remainder are listed in Table I. Only two English phonemes - /dh/ and /th/ - were not used.

This model was then used to align the lectures (again inserting a garbage model between words), followed by the DP alignment procedure described above and corpus segmentation. MAP adaptation was then performed on a per-lecture basis (where enough data was available), and globally for use with those lectures with less than 15 minutes of audio – the same training and test segments as for the NCHLT corpus were employed. The process of alignment and corpus segmentation was repeated using the MAP adapted model. The segmented corpus resulting from this second iteration was then used to train a new Afrikaans model from scratch, using the original Afrikaans phone set. These models were again MAP adapted on a per lecture basis.

3) Alignment durations: The direct alignment of long lectures can be computationally very expensive. In order to get an intuition of the relationship between audio length and alignment duration – and thus, to decide whether a pre-segmentation step is justified – we aligned various lengths of segmented lectures.

TABLE I
Source models for Afrikaans where direct mappings were not available.

English (ARPABET)	Afrikaans (SAMPA)	Example	
		English	Afrikaans
iy ax; ih ax	i@	peer	geen
uw ax; uh ax	u@	poor	boom
ih; iy	i	kin, keen	sien
iy ax; ih ax	2:	peer	museum
ax	9	(SAE) this	put
ih; iy	y	kin, keen	vuur
hh	x	hand	gaan
aw	a u	allow	gauteng
ay	a i	abide	baie
ch	t S	choke	tjalie
oy	O i	ahoy	boikot
jh	d Z	joke	jazz

IV. Analysis and Results

Several experiments were conducted to determine to what degree lectures can be processed in resource scarce environments. We discuss the implications of lecture duration and the corresponding computational cost of alignment in Section IV-B. Different approaches to improving alignment accuracy are discussed in Section IV-C and the implications of speaker adaptation on alignment accuracy is described in Section IV-D.

A. Experimental setup

From an acoustic modeling perspective, two different approaches were followed; a well trained Afrikaans model (described in section III-B1) was used to align Afrikaans lectures, and a bootstrapping approach was followed whereby a well trained American English model (see section III-B2) was used for alignment of the same set of lectures. The reason for the two different approaches is to firstly test the feasibility of lecture transcription in resource-scarce environments where no data in the target language is available, and secondly to determine how detrimental the lack of target-language acoustic models is at various stages in the processing chain.

In order to quantify the quality of our bootstrapped model, as well as alignment accuracy, a time aligned, accurate evaluation set is required. Since the audio in our corpus is only accompanied by approximate transcriptions, word or phone recognition – which would have been the standard way of evaluating our bootstrapped model – is infeasible. The transcriptions are also not time aligned, complicating the evaluation of our alignment accuracy.

Davel *et al.* [10] used 3 measures to quantify corpus and model improvement while processing internet harvested corpora accompanied by approximate transcriptions: (a) amount of audio absorbed by the garbage model, (b) average frame log likelihood and (c) average DP score. These measures were found to correlate well with each other, as well as with

phoneme accuracy as estimated on a carefully transcribed subset of data. We adopted these same measures to quantify performance on spoken lecture alignment.

Another measure, duration-independent overlap rate [18] was employed to evaluate alignment accuracy. For this measure, 100 randomly selected word instances were manually time aligned across the corpus (50 within the segments that were used for MAP adaptation, and 50 in the segments that were used for evaluation only).

Eight systems were evaluated. The following notation is employed in tables II and III:

- **WSJ** refers to the WSJ model where phone mappings, as described in table I were employed. A garbage model was also used in conjunction with this model.
- **WSJ no gm** is the same as the model mentioned above, except that no garbage model was trained. This is the only instance of the WSJ process where no garbage modeling was used.
- **WSJ + MAP/spk** refers to speaker adaption using the training subset from the longer lectures, on a per lecture basis. A pooled model was also created by performing MAP adaptation using all training data (including the shorter lectures). Whenever a lecture had its own MAP adapted model, that model was then used to align and segment the lecture, using techniques described in [10]. If no such model exists, the pooled model was used for segmentation.
- The segmented "corpus" of training data was then used to retrain an Afrikaans spoken lecture (**ASL**) model from scratch. Since the segmented corpus contains Afrikaans lecture transcription data, the original Afrikaans phoneset was used. This model was again used to segment the corpus.
- **ASL + MAP/spk** refers to another round of speaker adaptation using the ASL model
- For the assumption of the existence of a target-language acoustic model, we used the Afrikaans NCHLT corpus (**NCHLT**), together with a garbage model.
- **NCHLT no gm** is again the only instance where no garbage modeling was used.
- Speaker adaptation was again used to adapt this model on a per lecture basis (**NCHLT + MAP/spk**).

B. Time implications of alignment length

University lectures in South Africa are typically at least 30 minutes or longer in duration. As mentioned before, alignment of such lectures is useful for indexing and general processing. Naive alignment is expected to take very long, though, as other researchers have noted [19]. However, solutions in resource rich languages entail the use of language models. In the absence of well-trained language models, the computational cost of aligning very long audio segments and the corresponding trade-off with manual labour becomes an important issue to investigate.

For this purpose, we manually segmented the spoken-lectures corpus into 5, 10, 20 minute and longer (original duration) segments. Alignment and decoding of each segment were performed and timed on a single processor of a standard desktop compute server. Figure 1 displays the result of this timed alignment (and also for decoding, which we discuss below). A relationship that is significantly worse than linear is observed between alignment duration and the duration of the audio being aligned. The longest lecture, which is 42 minutes in duration, took approximately 97 hours to align, whereas the short segments were aligned in less than real time.

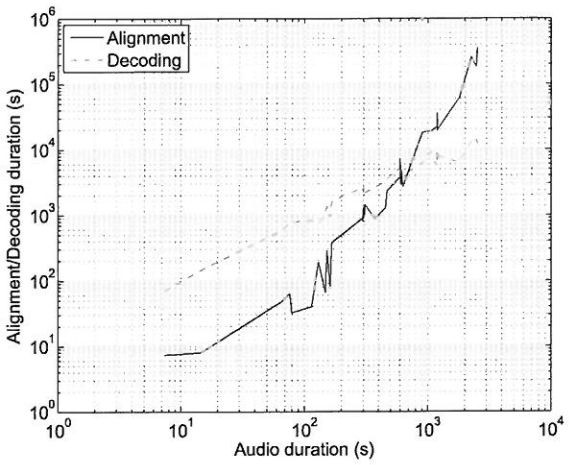

Fig. 1. Alignment and decoding times as a function of audio duration, for several lectures and lecture segments.

The graph confirms that either manual intervention, where lectures are segmented into 5 minute segments, or some automatic approach where smaller chunks are iteratively aligned against the text until the best match is found, would be much more computationally efficient than simply aligning entire lectures.

C. Alignment accuracy

Alignment accuracy was evaluated using 4 measures, as described in Section IV-A. Duration independent overlap was measured across two sets of manually time aligned words. The first set of 50 words was selected from those lectures which are long enough to allow for a sizable amount of speech for MAP adaptation (after some held-out sections have been removed, typically the last half of the lecture). Another set of 50 words was selected from the remaining, much shorter lectures. DP scores, average frame log likelihood and percentage of audio absorbed by the garbage model were only measured on the held-out subset from the longer lectures.

From Tables II and III, it is clear that having a target-language acoustic model available is the best case scenario. However, it is very encouraging to see that a bootstrapping approach can get very similar performance, at least as far as alignment accuracy is concerned. It would be interesting to see how these results scale when performing indexing or word recognition.

TABLE II
Duration independent overlap rate when using different models for alignment.

model	50 words	100 words
WSJ	66.94	61.62
WSJ + MAP/spk	77.79	-
WSJ no gm	80.54	73.63
NCHLT no gm	83.65	76.84
ASL	85.80	-
ASL + MAP/spk	86.64	-
NCHLT	88.53	82.76
NCHLT + MAP/spk	93.37	-

TABLE III
Improvements observed during model refinement and alignment, reported on the evaluation set.

model	Avg DPS	log P	non-speech (%)
WSJ no gm	-0.270	-91.79	36.37
WSJ	-0.130	-88.46	58.60
NCHLT no gm	-0.076	-84.28	27.68
WSJ + MAP/spk	-0.063	-84.56	45.74
NCHLT	-0.016	-84.05	46.29
ASL	0.163	-77.46	42.17
ASL + MAP/spk	0.217	-76.77	41.67
NCHLT + MAP/spk	0.236	-81.42	42.53

D. The effect of speaker adaptation

As expected, speaker adaptation seems to be beneficial to the process of alignment. This is most obvious from the difference between the system trained on segmented ASL data, compared to the ASL system which was MAP adapted to particular speakers.

The spoken lecture domain is different from general ASR corpora, in that one can expect to receive substantial amounts of data from a particular speaker who is lecturing a course. Our hope is thus that as the corpus grows, enough data will become available to train speaker specific models instead of having to adapt from a larger model. The amount of data at which this transition occurs (i.e. where a speaker-specific model is more accurate than a general, adapted model) will be investigated in future work.

E. The effect of garbage modeling

The garbage model we employed was very effective when used with the NCHLT model, as can be seen from Table III. All measurements (except the percentage of non-speech audio, which is not easy to interpret without knowledge of how much real speech is present) agree that garbage modeling is very beneficial to the process.

On the WSJ model, the picture looks quite different, though. From the amount of non-speech, it seems that our garbage model is too greedy. The exact reason why it misbehaves with WSJ as opposed to NCHLT is an interesting and important research question that needs to be answered for this technique to be completely robust. Our hypothesis is that language differences (between the model being employed and the audio being aligned) can lead to predictable failures of the garbage-model approach. In particular, if the triphone models do not model the target language well, the more general garbage model (which has a large variance) becomes a better match than the closest matching phone. One easy remedy may be to use fewer mixtures with the triphone models, or to explicitly penalize the garbage model based on language distance measures, such as the Levenstein distance.

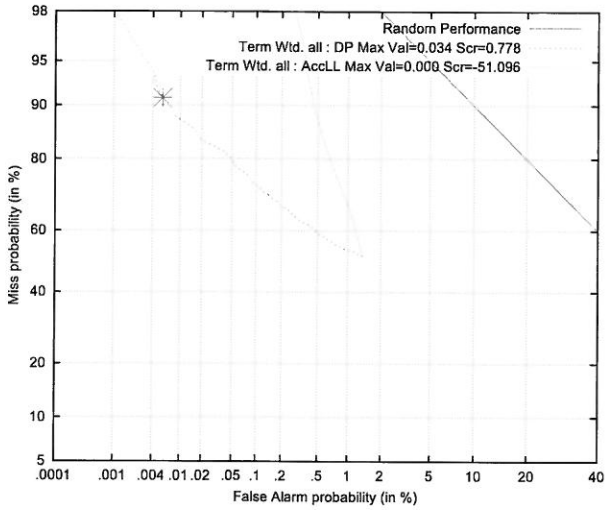

Fig. 2. DET curves displaying the result of an initial STD system. Results are displayed when using average frame log-likelihoods and DP scores respectively.

F. Initial indexing results

Initial spoken term detection (STD, or indexing) results are displayed in figure 2. STD was performed by using a grammar which allows any number (including zero) of the specific term being searched for to be detected in a single file. As an alternative, a garbage-silence hybrid model is allowed in parallel and between terms.

DP scoring using a flat confidence matrix, as used in [10], is clearly the superior confidence measure when compared to average frame log-likelihoods. Future work will entail evaluating indexing on accurately transcribed lectures and using either linguistically motivated or data-driven scoring matrices for DP scoring.

V. Conclusion

Initial work towards processing spoken lectures in resource-scarce environments has been presented. It has been shown that having target-language acoustic models is beneficial to the processing of these lectures in general. However, the results from using a well-trained model from a different language in a recursive bootstrapping procedure are encouraging. We will in future evaluate how these results correlate with actual lecture transcription word error rates. The WSJ results also seem to suggest (in line with observations in [10]) that with only a couple of hours of approximately transcribed lectures, one should be able to erase the initial difference observed between

starting the process with a target-language model, as opposed to a model from a different language. While we expect this technique to generalize well to other resource-scarce languages where reasonable phone mappings can be found, it will be interesting to verify.

It will also be interesting to see to what extent our initial corpus can be extended, using a recursive word-recognition and unsupervised-retraining approach. Given the results that we have achieved to date, we are confident that lecture-transcription services with a usable accuracy can be created for languages with limited resources – we therefore also look forward to investigating the practical application of such services in multilingual schools and universities.

References

[1] K. Bain, S. Basson, and M. Wald, "Speech Recognition in University Classrooms: Liberated Learning Project," in *Proceedings of the fifth international ACM conference on Assistive technologies*. ACM, 2002, pp. 192–196.

[2] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld, "Healthline: Speech-based Access to Health Information by low-literate users," in *Proc. IEEE Int. Conf. on ICTD*, Bangalore, India, Dec. 2007, pp. 131–139.

[3] M. Plauché and U. Nallasamy, "Speech interfaces for equitable access to information technology," *Information Technologies and International Development*, vol. 4, no. 1, pp. 69–86, 2007.

[4] E. Barnard, M. Davel, and G. van Huyssteen, "Speech Technology for Information Access: a South African case study," in *AAAI symposium on AI for D*, Palo Alto, CA, March 2010, pp. 8–13.

[5] J. Sherwani, "Speech interface for information access by low-literate users in the developing world," Ph.D. dissertation, Computer Science Department, Carnegie Mellon University, 2009.

[6] C. van Heerden, E. Barnard, and M. Davel, "Basic speech recognition for spoken dialogues," in *Proc. Interspeech*, Brighton, UK, Sept. 2009, pp. 3003–3006.

[7] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech*, Antwerp, Belgium, Sept. 2007, pp. 2553–2556.

[8] T. Kawahara, "Automatic transcription of parliamentary meetings and classroom lectures - A sustainable approach and real system evaluations," in *Proceedings ICSCLP*, Tainan, Taiwan, November 2010.

[9] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[10] M. H. Davel, C. van Heerden1, N. Kleynhans, and E. Barnard, "Efficient harvesting of Internet audio for resource-scarce ASR," in *Proceedings Interspeech*, Florence, Italy, August 2011, pp. 3153–3156.

[11] N. T. Vu, F. Kraus, and T. Schultz, "Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training," in *Proc. Interspeech*, Florence, Italy, month = Sep., year = 2011, pages = 3145–3148.

[12] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Proc. Interspeech*, Florence, Italy, month = Sep., year = 2011, pages = 1693–1696.

[13] S. Chaudhuri, M. Harvilla, and B. Raj, "Unsupervised Learning of Acoustic Unit Descriptors for Audio Content Representation and Classification," in *Proc. Interspeech*, Florence, Italy, month = Sep., year = 2011, pages = 2265–2268.

[14] M. Davel and E. Barnard, "Pronunciation predication with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, Oct. 2008.

[15] M. Davel and F. de Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010, pp. 1898–1901.

[16] N. J. de Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. de Waal, "Woefzela - An open-source platform for ASR data collection in the developing world," in *Proceedings Interspeech*, Florence, Italy, August 2011, pp. 3176–3179.

[17] S. Young, G. Evermann, M. Gaels, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4," March 2009.

[18] S. Paulo and L. Oliveira, *Advances in Natural Language Processing*. Berlin / Heidelberg: Springer, 2004.

[19] P. J. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Proc. ICSLP*, 1998, paper 0068.