THE ZERO RESOURCE SPEECH CHALLENGE 2017
Ewan Dunbar, Xuan Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, Emmanuel Dupoux

To cite this version:
Ewan Dunbar, Xuan Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, et al.. THE ZERO RESOURCE SPEECH CHALLENGE 2017. IEEE Automatic Speech Recognition and Understanding (ASRU), Dec 2017, Okinawa, Japan. <hal-01664586>

HAL Id: hal-01664586
https://hal.archives-ouvertes.fr/hal-01664586
Submitted on 15 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT

We describe a new challenge aimed at discovering subword and word units from raw speech. This challenge is the followup to the Zero Resource Speech Challenge 2015. It aims at constructing systems that generalize across languages and adapt to new speakers. The design features and evaluation metrics of the challenge are presented and the results of seventeen models are discussed.

Index Terms— zero resource speech technology, subword modeling, acoustic unit discovery, unsupervised term discovery

1. INTRODUCTION

Traditional speech technology relies on linguistic expertise and textual information to build acoustic and language models. Recent work has begun to mitigate the need for expert knowledge, having succeeded at training supervised systems using textual transcripts only [1, 2]. But for the majority of the world’s languages, even textual transcripts are sparse or non-existent, and for many languages, the lack of standard orthography would reduce the utility of such transcripts greatly. This is why there is increasing interest in a radically different approach, the so-called zero resource setting [3, 4]. The aim is to construct speech systems without any textual or linguistic resource. The fact that young children become fluent communicators in their native language(s) before they learn to read and write, and with minimal or no expert supervision, demonstrates that such a system is possible.

Systems constructed with zero expert resources have at least three kinds of applications. They can form the basis for speech services for the perhaps billions of speakers of languages without readily available linguistic or textual resources in sufficient quantities. They can also help field linguists document endangered languages, by providing tools to semi-automatically analyze and annotate audio recordings using automatically discovered linguistic units (phonemes, lexicon, grammar) [5]. In this case, a realistic scenario is to process raw speech and its translation in a well-resourced language [6, 7, 8]. Finally, to the extent that they mimic language development, these models can be used to evaluate scientific claims about development, and act as quantitative tools for psychologists and clinicians interested in the impact of sociolinguistic variations in input on subsequent normal or abnormal language and cognitive development [9].

As the previous Zero Resource Speech Challenge [10, 11], the 2017 installment consists of two different sub-challenges. Participants can submit systems that do either subword modelling (track one) or spoken term discovery (track two). Subword modelling means constructing a representation of speech sounds; we say “subword” instead of “phoneme” or “sound” because the criterion for success does not demand any particular temporal window size. The requirement is to have a representation of the speech signal that is robust to talker variation, and makes the distinctions necessary to distinguish words. Spoken term discovery means discovering recurring speech fragments in the audio: the system must output time-stamps delimiting stretches of speech, associated with class labels, ideally corresponding to real words in the language. The systems must be able to do their task without linguistic resources. The 2017 Zero Resource Speech Challenge [1] has two main innovations compared to the 2015 iteration: (1) it tests how well systems’ architectures and hyperparameters generalize to unseen languages (the systems are developed with three languages and tested on two ‘surprise’ languages); (2) it tests how well the trained systems’ parameters adapt to unseen talkers (varying the amount of speech available for each talker).

2. THE CHALLENGE

2.1. Track one: Subword modelling

The goal of subword modelling is to learn representations appropriate to the sounds of the language (phones or phonemes). As adults, we easily perceive differences between individual phonemes in our native language, substantially better than in non-native languages [12]. This so-called perceptual tuning develops substantially during the first year of life [13, 14].

1http://zerospeech.com/2017
The input to subword modelling is a set of acoustic features encoding continuous speech. The result is a transformation that applies to new samples of running speech, yielding a new representation (continuous or discrete) of that speech. After training in a particular language from unlabelled speech, the representation is computed for a test set in the same language on novel speakers, to be evaluated on phoneme discrimination based on gold-standard annotations.

2.2. Track two: Spoken term discovery

Intuitively, the goal of spoken term discovery is to find words in the speech stream—just as the infant learns the words of its language by listening. The input to candidate systems is a series of speech features. The output is a set of boundaries defining the start and end of proposed word tokens discovered in the speech, and category labels indicating proposed word types. These boundaries may, but need not, constitute an exhaustive parse of the speech. The evaluation we apply is a set of scores measuring different aspects of the alignment with the words in the gold-standard transcription. As is customary in the field of word segmentation, we do not provide a separate test set for this track; we rely on the surprise languages to assess possible hyperparameter overfitting.

2.3. Generalization across languages

This challenge is explicitly set up as a ‘learning to learn’ or meta-learning problem: participants select the best architectures and tune their hyperparameters on three development languages on which the evaluation software is provided for a test set. The higher level test is performed on two new surprise languages for which the evaluation is withheld. Participants submit the output of their previously optimized systems for these two corpora, and the organizers perform the evaluation. Each submission (maximum five per team) is automatically evaluated and made public.

2.4. Adaptation to new speakers

The subword modeling task is explicitly set up as a talker adaptation problem. The training data is constructed from a skewed distribution of speakers, and the test set consists only of new speakers appearing in files of various durations (from 1 second to two minutes). This enables us to evaluate models’ ability to generalize and adapt to new speakers on the basis of limited evidence.

3. DATA SETS

Two groups of data sets are provided as part of the challenge: the development data and the surprise data. The development data consists of corpora from three different languages (English, French and Mandarin). Each corpus comes with software that does the evaluation for the two tracks. Challenge participants are encouraged to use these resources to tune their hyperparameters using a cross validation approach to maximize generalizability. Each corpus is split into a train and a test set; the subword modelling evaluations are run on the test sets, and the spoken term discovery evaluations on the training sets. The participants then have to submit their systems and their output on all the data sets for independent evaluation (run automatically upon submission). The surprise data consists of corpora from two new languages, the identity of which was not revealed to participants. Participants receive only the speech corpora, with no additional resources. As before, each contains a training set (which participants can use to train their systems) and a test set for track one evaluation.

The amount of data in the training part of the development data sets varies from 2.5 to 45 hours (see Table 1), to ensure that systems can work both with limited data and with large data sets. The statistics of the two surprise languages fall between these two extremes. The distribution of speakers in the training sets is shaped to reflect what is typically found in natural language acquisition settings: there is a “family”—a small number of speakers that make up a large proportion of the total speech—and a set of “outsiders”—a larger number of speakers that each appear in smaller proportions. The test sets consist of many short files, are organized into subsets of differing length (1s, 10s and 120s).

The English and French corpora were taken from the LibriVox collection of audio books and phone force-aligned using Kaldi [15]. The Mandarin corpus is the one described in [16], force-aligned using Kaldi. The first surprise language is German, and was taken from LibriVox and force-aligned using Kaldi as well. The second surprise language is Wolof, and is the corpus described in [17].

4. EVALUATION

4.1. Evaluation of subword modelling

Unsupervised subword modeling is a representation learning problem: the system must find speech features that emphasize linguistically relevant properties (phoneme structure) and de-emphasize the linguistically irrelevant ones (speaker identity, emotion, channel, etc). We use a minimal pair ABX task [18] [19], which does not require any training, and only requires a dissimilarity metric between speech tokens. The ABX task is inspired by match-to-sample tasks used in human psychophysics and is a simple way to measure discriminability between two sound categories (where the sounds A and B belong to different categories x and y, respectively, and the task is to decide whether the sound X belongs to one or the other). We define the ABX-discriminability of category x from category y as the probability that A and X are further apart than B and X when A and X are from category x and B

\[ P(X_A > X_B | A = x) \]

\[ P(X_B > X_A | B = y) \]

\[ P(X_A > X_B | A = x, B = y) \]

\[ P(X_B > X_A | B = y, A = x) \]
Table 1. Corpus statistics

<table>
<thead>
<tr>
<th>Development</th>
<th>English</th>
<th>French</th>
<th>Mandarin</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>#speakers</td>
<td>9</td>
<td>10</td>
<td>4</td>
<td>S1 – German, S2 – Wolof</td>
</tr>
<tr>
<td>duration/ speaker</td>
<td>165-220 min</td>
<td>110-195 min</td>
<td>20-25min</td>
<td>10-85-150 min</td>
</tr>
<tr>
<td>#words</td>
<td>45h</td>
<td>24h</td>
<td>2 min</td>
<td>25h</td>
</tr>
<tr>
<td>#files</td>
<td>370k</td>
<td>220k</td>
<td>20k</td>
<td>2383</td>
</tr>
<tr>
<td>duration (min)</td>
<td>370k</td>
<td>220k</td>
<td>20k</td>
<td>2383</td>
</tr>
</tbody>
</table>

Table 2. Track 1 within and between talker ABX scores for three test file durations, for the development (average) and surprise languages, for the 14 submitted systems (the mean performance on the surprise set is in square brackets). The three supervised systems are listed separately. See [http://zerospeech.com/2017](http://zerospeech.com/2017) for the full table including development languages.

<table>
<thead>
<tr>
<th></th>
<th>Avg. Development</th>
<th>S1 – German</th>
<th>S2 – Wolof</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [18.84]</td>
<td>12 12.1 12.1</td>
<td>14.2 13.3 12.9</td>
<td>14.1 14.3 14.1</td>
</tr>
<tr>
<td>Topline [7.19]</td>
<td>8 5.4 5.3</td>
<td>8.7 7.1 7.0</td>
<td>8.6 6.6 6.4</td>
</tr>
<tr>
<td>H [9.16]</td>
<td>8.5 7.6 7.4</td>
<td>8.5 5.6 5.3</td>
<td>10.9 8.8 8.4</td>
</tr>
<tr>
<td>P1 [14.90]</td>
<td>10.7 9.3 8.7</td>
<td>7.9 6.9 6.3</td>
<td>12.7 11.9 11.3</td>
</tr>
<tr>
<td>P2 [14.9]</td>
<td>10.8 8.7 8.5</td>
<td>8.8 6.6 6.3</td>
<td>13.1 11.7 11.7</td>
</tr>
<tr>
<td>C1 [11.95]</td>
<td>10.0 8.4 8.3</td>
<td>7.6 6.2 6.3</td>
<td>11.7 9.9 9.8</td>
</tr>
<tr>
<td>C2 [11.24]</td>
<td>10.1 8.5 8.4</td>
<td>7.6 6.2 6.1</td>
<td>11.6 9.8 9.6</td>
</tr>
<tr>
<td>A1 [12.3]</td>
<td>8.8 7.9 7.8</td>
<td>6.9 6.1 6.0</td>
<td>9.9 9.2 9.1</td>
</tr>
</tbody>
</table>

Task, A and B belong to the same speaker, and X to a different one. A = beg₁ T₁, B = bag₂ T₁, X = bag₂ T₂. The scores for a given minimal pair are first averaged across all of the pairs of speakers for which this contrast can be made. As above, the resulting scores are averaged over all contexts over all pairs of central phones and converted to an error rate.

4.2. Evaluation of spoken term discovery

“Spoken term discovery” is a broad label. We use it to refer to systems that do one or more of three tasks. Systems that do matching find fragments that are matched pairwise as being instances of the same sequence of phonemes, attempting to find as many as possible. They can be evaluated based on how similar the matched fragments are, and how much of the corpus they cover. Systems that do lexicon discovery group these fragments into sets (rather than just matching pairwise), with the goal of finding a high-quality lexicon of types. They can be evaluated based on how well the sets match on the sequence of phonemes, and how well the sets match the gold-standard lexicon of word types. Systems that do word segmentation attempt to find fragments that are aligned with the gold-standard word transcription. By setting out three different types of criteria, the intention is to be open to various types of “spoken term discovery” systems, all of which in some sense “find words.” The result is that we do three

is from category y, according to a dissimilarity function d. We obtain a symmetric discriminability score by taking the average of the ABX discriminability of x from y and of y from x. The dissimilarities provided in this challenge are based on dynamic time warping, the underlying frame-level dissimilarity being either the cosine distance or KL-divergence. For most systems (signal processing, embeddings) the cosine distance usually gives good results, and for others (posteriorigrams) the KL divergence is more appropriate. Participants are allowed to supply their own dissimilarity as long as it was not obtained through supervised training.

We focus on phone triplet minimal pairs: sequences of 3 phonemes that differ in the central sound (not necessarily real words, e.g., “beg”–“bag”, “api”–“ati”, etc). Our compound measure sums over all minimal pairs of this type found in the corpus in a structured manner, that depends on the task. For the within-speaker task, all of the phone triplets belong to the same speaker (e.g. A = beg₁ T₁, B = bag₂ T₁, X = bag₂ T₂). The scores for a given minimal pair are first averaged across all of the speakers for which this minimal pair exists. The resulting scores are then averaged over all found contexts for a given pair of central phones (e.g. for the pair /l/-/l/, average the scores for the existing contexts such as /b,/, /l,/, /l/s/, etc.). Finally the scores for every pair of central phones are averaged and subtracted from 1 to yield the reported within-talker ABX error rate. For the across-speaker
4.2.1. Evaluation as matching system

The evaluation of spoken term discovery systems as matching systems consists of two scores, NED and coverage. NED, is the average, over all matched pairs, of the Levenshtein distance between their phonemic transcriptions, divided by the max of their phonemic length. The second score is the coverage, the fraction of the discoverable part of the corpus that is covered by all the discovered fragments. The discoverable part of the corpus is found by computing the union of all of the intervals corresponding to all of the pairs of ngrams (with n between 3 and 20). This is almost all of the corpus, except for unigram and bigram hapaxes.

4.2.2. Evaluation as lexicon discovery system

Six scores are used to evaluate the performance of a spoken term discovery system in terms of lexicon discovery. The first three are grouping precision, recall and F-score. These are defined in terms of \( P_{clus} \), the set of all pairs of fragments that are groups in the same cluster, and \( P_{goldclus} \), the set of all non-overlapping pairs of fragments which are both discovered by the system (not necessarily in the same cluster) and have exactly the same gold transcription.

\[
\begin{align*}
\text{Prec:} & \quad \sum_{t \in \text{types}(P_{clus})} w(t, P_{clus}) \frac{\text{occ}(t, P_{clus} \cap P_{goldclus})}{\text{occ}(t, P_{clus})} \\
\text{Rec:} & \quad \sum_{t \in \text{types}(P_{goldclus})} w(t, P_{goldclus}) \frac{\text{occ}(t, P_{clus} \cap P_{goldclus})}{\text{occ}(t, P_{goldclus})}
\end{align*}
\]

Where \( t \) ranges over the types of fragments (defined by the transcription) in a cluster \( C \), and \( \text{occ}(t, C) \) is the number of occurrences of that type, \( w \) the number of occurrences divided by the size of the cluster. In other words, Prec is a weighted measure of cluster purity and Rec, of the inverse of the cluster’s fragmentation. The other three scores are type precision, recall, and F-score. Type precision is the probability that discovered types belong to the gold set of types (real words), whereas type recall is the probability that gold types are discovered. We restrict both sets to words between three and twenty segments long.

4.2.3. Evaluation as word segmentation system

Six scores are used to evaluate the performance of a spoken term discovery system as a system for word segmentation. The first three are token precision, recall, and F-score. Token precision is the probability that discovered fragment tokens are in the gold set of word fragments, and token recall the probability that the gold fragments are discovered. Again, tokens are restricted to be between three and twenty segments long. Finally, boundary precision, recall, and F-score are similarly defined with respect to the set of discovered and gold word boundaries.

5. DESCRIPTION OF SUBMITTED SYSTEMS

We received sixteen systems for track one, and three for track two. For two track one submissions, the authors declared that the systems were not intended as submissions, and are thus not included in the present overview, although the results can be found on the leaderboard on the Challenge website [zerospeech.com/2017].

5.1. Track 1

The baseline features are the MFCC features provided to participants; they are evaluated using the ABX score using the cosine similarity. The topline system is a supervised HMM-GMM phone recognition system, with a bigram language model, trained with a Kaldi recipe [15]. The phone posteriors are evaluated using the KL divergence.

The submitted systems follow four strategies. The first strategy consists in performing a bottom-up frame-level clustering, following up on the success of [20] with this simple strategy. In system [H], from Heck et al. [21], PLP representations are used, as well as the output of three learned feature transformations, to be the input to the same frame-level clustering algorithm used in [20] (the parallel Dirichlet process Gaussian mixture model sampler of [22]). The transformations (LDA, LDA+MLLT; LDA+MLLT+fMLLR) are aimed at making the representation talker invariant. The training requires labels, which are obtained from an initial clustering using the DP-GMM algorithm. The submitted system is a combination of the raw posteriorgrams (not smoothed by any language model) from the four DP-GMMs obtained by clustering frames in each of the three transformed representations, plus the raw PLP features. System [P1], from Pellegrini et al. [23], uses \( k \)-means clustering on ZCA-transformed (whitened using PCA and rotated back to the original axes) MFCC features. In system [P2], this is followed by a step of centroid re-estimation: the centroids are re-estimated based only on the data points for which the label is the same as that of its left and right neighbors. The distance from point to all cluster centroids is the feature vector in both.

A second strategy constructs language-independent embeddings by training neural networks in an unsupervised way. The systems of Chen et al. [24] first cluster frames separately on each language, using the same DP-GMM algorithm as [21] and [20]. The cluster labels are then used as tar-
Table 3. Track 2 results. Scores are averages over evaluations done on subparts of the corpus: the list of corpus gold utterances is split in groups of 1000 elements/utterances in order to compute the scores in parallel. Thus, the F-scores presented are not directly related to the precision and recall scores.

<table>
<thead>
<tr>
<th>Model</th>
<th>Words</th>
<th>Pairs</th>
<th>NED</th>
<th>smoothing</th>
<th>Grouping type</th>
<th>Token type</th>
<th>Boundary</th>
<th>Avg. Training</th>
<th>English, French, Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>4948</td>
<td>5842</td>
<td>30.0</td>
<td>4.1</td>
<td>54.8</td>
<td>48.7</td>
<td>52.3</td>
<td>5.5</td>
<td>0.3</td>
</tr>
<tr>
<td>K</td>
<td>24724</td>
<td>375685</td>
<td>76.0</td>
<td>105.0</td>
<td>6.7</td>
<td>8.3</td>
<td>6.7</td>
<td>4.6</td>
<td>9.0</td>
</tr>
<tr>
<td>G1</td>
<td>51377</td>
<td>465997</td>
<td>71.5</td>
<td>60.5</td>
<td>3.9</td>
<td>67.8</td>
<td>7.4</td>
<td>3.9</td>
<td>6.0</td>
</tr>
<tr>
<td>G2</td>
<td>457369</td>
<td>71.4</td>
<td>60.4</td>
<td>4.0</td>
<td>67.6</td>
<td>7.5</td>
<td>3.9</td>
<td>6.0</td>
<td>4.4</td>
</tr>
<tr>
<td>S1 – German</td>
<td>2973</td>
<td>3315</td>
<td>80.5</td>
<td>3.0</td>
<td>54.8</td>
<td>94.6</td>
<td>64.9</td>
<td>5.5</td>
<td>0.3</td>
</tr>
<tr>
<td>K</td>
<td>28675</td>
<td>425873</td>
<td>66.4</td>
<td>100.0</td>
<td>11.8</td>
<td>5.9</td>
<td>7.9</td>
<td>5.7</td>
<td>11.2</td>
</tr>
<tr>
<td>G1</td>
<td>60848</td>
<td>58209</td>
<td>59.9</td>
<td>71.8</td>
<td>5.7</td>
<td>63.8</td>
<td>10.4</td>
<td>3.2</td>
<td>6.8</td>
</tr>
<tr>
<td>G2</td>
<td>60489</td>
<td>588162</td>
<td>59.9</td>
<td>71.8</td>
<td>5.8</td>
<td>64.0</td>
<td>10.7</td>
<td>3.2</td>
<td>6.8</td>
</tr>
</tbody>
</table>

5.2. Track 2

The baseline of track two was computed using [30], which does pair-matching using locally sensitive hashing applied to PLP features and then groups pairs using graph clustering. The parameters of the STD stayed the same across all languages, except that the DTW threshold was increased for Mandarin (to 0.90, from 0.88) in order to obtain a NED value similar to that of other languages. The topline was an exhaustive-parsing word segmentation model based on the textual transcriptions (a unigram grammar trained in the adaptor grammar framework: [32]).

System [K], from Kamper et al. [33], uses k-means to discover recurring acoustic patterns, jointly optimized with an exhaustive segmentation. The variable-length acoustic chunks of speech yielded by a segmentation are reduced to fixed dimension by down-sampling in the time dimension for clustering by k-means. The systems of García-Granada...
et al. use a supervised ASR system for Hungarian to decode the speech. Chunks of (transcribed) speech match if they have the same transcription. These matches are then filtered: a representation of the speech is obtained by training an autoencoder, and only pairs with DTW sufficiently low in this representational space (below a threshold) are retained (system [G1]). In system [G2], pairs are tolerated that have a one-phone difference in their transcriptions.

6. RESULTS

The results of track one are summarized in Table 2. The clear winner among the unsupervised models is [H]. While the paper focuses on the role of learned feature transformations—and demonstrates that they improve the evaluation—these are unlikely to be solely responsible for the very good performance of the system. The baseline system described in the paper, with no transformations applied, consisting of posteriorgrams from a frame-wise DPGMM clustering, still performs extremely well. Although it was not evaluated on the surprise languages, its average ABX score on the development languages (9.7) is still better that of the second-ranked system [C2] (10.7), from Chen et al. The systems presented by Heck et al. all have in common monolingual training, direct evaluation of GMM posteriorgrams, and the fact that they use PLP features as input, rather than MFCCs.

The systems of Chen et al. (as well as Yuan et al.) use an initial stage of monolingual training, followed by multilingual training. Among the other systems in [24] (not submitted for final evaluation) are some with fully monolingual training, which serve to partly evaluate the effect of multilingual training: monolingual training fares better, within language. However, as the numerical scores are not presented, it is not possible to determine whether the use of monolingual training could account for the good scores of Heck et al. What is clear is that multilingual training still works reasonably well—all the remaining systems, except for the lower performing system [P], use multilingual training, and show substantial improvements over the baseline. The setting for these systems corresponds to the situation of an infant raised in a multilingual environment.

The STD strategy for improving subword features fares relatively poorly; a limitation of the Yuan et al. systems is that they use an architecture (autoencoders) that was previously reported to be suboptimal for improving acoustic features [28]. A Siamese network [27] might improve the result.

The results of track two are in Table 3. As in the previous challenge, the baseline provided by the system of [30] has a high NED (very good matching), but has a low coverage (few matches). Given that system [K] demands an exhaustive parse, it is guaranteed to have full coverage, which risks poorer matches. It is thus notable that system [K] has a NED on the development languages comparable to the partly supervised systems [G1] and [G2]. On the other hand, grouping by supervised transcription clearly works better than k-means clustering; while the groups have low precision (many erroneously grouped items) for both kinds of systems, [G1]/[G2] have much higher grouping recall (more of the true group members among the items grouped together) than system [K]. The quality of segmentation, in terms of tokens and of boundaries, is, on the other hand, much better in system [K], presumably as a result of increased coverage.

7. CONCLUSION

The results of the 2017 Zero Resource Speech Challenge build very clearly on the previous challenge. The strategies most successful in the previous challenge for track one, bottom-up clustering and training based on pairs of matching words discovered using spoken term discovery—are built upon, and a new strategy, multilingual training, is introduced. This strategy, although not completely foreseen by the organizers, would be similar to the situation of a multilingual infant. This new approach is used in conjunction with additional tools, including further training using neural networks and the application of time series models. Interestingly, simple bottom-up clustering still seems to work best in the unsupervised case. The fact that supervised labels trained fully out of domain seem to yield good results in both track one and track two (on Japanese, in the case of [S1], and on Hungarian in the case of [G1] and [G2]) indicates that the approach of having systems with strong prior knowledge about human speech is still a benchmark to be beaten in the unsupervised case. (The excellent results of [S2] may yet be partly attributable to the use of supervised training in German, which was one of the surprise languages.) Further work is needed to systematically evaluate the effect of each of these different manipulations, however, and to directly compare these new approaches with those from the previous challenge.

Potential future lines of inquiry building on the new track two systems might include exploring using these systems to feed subword modelling, especially in combination with some of the new subword modelling strategies used here, in light of the successful use of STD to improve subword models in the previous challenge, and the reasonable improvement over baseline shown in this challenge. In the other direction, the major unsupervised improvements to the speech features demonstrated in track one suggest strong room for improvement in unsupervised spoken term discovery systems like [K].

The 2017 ZeroSpeech Challenge remains open for submissions exploring these, or any other new strategies for discovering phoneme- and word-like units without access to linguistic resources.
8. REFERENCES


