Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking rosetta" JSALT 2017 workshop

Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stuker, Pierre Godard, Markus Muller, Lucas Ondel, et al.

LINGUISTIC UNIT DISCOVERY FROM MULTI-MODAL INPUTS IN UNWRITTEN LANGUAGES: SUMMARY OF THE "SPEAKING ROSETTA" JSALT 2017 WORKSHOP

Odette Scharenborg[1*], *Laurent Besacier*[2], *Alan Black*[3], *Mark Hasegawa-Johnson*[4], *Florian Metze*[3], *Graham Neubig*[3], *Sebastian Stüker*[5], *Pierre Godard*[6], *Markus Müller*[5], *Lucas Ondel*[8], *Shruti Palaskar*[3], *Philip Arthur*[3], *Francesco Ciannella*[3], *Mingxing Du*[7], *Elin Larsen*[7], *Danny Merkx*[1], *Rachid Riad*[7], *Liming Wang*[4], *Emmanuel Dupoux*[7 †]

[1] Radboud University, [2] LIG - Univ Grenoble Alpes (UGA), [3] Carnegie Mellon University, [4] University of Illinois, [5] Karlsruhe Institute of Technology, [6] LIMSI CNRS, [7] ENS/CNRS/EHESS/INRIA, [8] Brno University.

ABSTRACT

We summarize the accomplishments of a multi-disciplinary workshop exploring the computational and scientific issues surrounding the discovery of linguistic units (subwords and words) in a language without orthography. We study the replacement of orthographic transcriptions by images and/or translated text in a well-resourced language to help unsupervised discovery from raw speech.

Index Terms— unwritten languages, multi-modal data, unsupervised unit discovery, image retrieval, machine translation.

1. INTRODUCTION

To develop speech and language technology (SLT) large amounts of annotated data are required. However, for many languages in the world, not enough speech data is available, or these lack the annotations needed to train an ASR system [1]. Moreover, an estimated half of the human languages do not have an orthography, and many others do not use it in a consistent fashion. This represents millions of potential users that as yet cannot be served by speech technologies. As any human 4-year-old demonstrates, however, it is theoretically possible to learn a language communication system before learning to read and write, from raw sensory signals and with only limited human supervision.

Recently, different approaches have been proposed to build ASR systems for such low-resource languages. One strand of research focuses on discovering the linguistic units of the low-resource language from the raw speech data, while assuming no other information about the language is available, and using these to build ASR systems (zero resource approach; e.g., [2, 3, 4, 5, 6]). Another strand of research focuses on building ASR systems using speech data from multiple languages, thus trying to create universal or cross-lingual ASR systems [7, 8, 9, 10]. Children though, when learning a language, also have information besides the auditory input available, primarily in the visual modality. This has led to a new strand of research which uses visual information, from images, to discover word-like units from the speech signal using speech-image associations [11, 12, 13]. The "Speaking Rosetta" project at the 2017 Frederick Jelinek Memorial Summer Workshop, which took place at Carnegie Mellon University, Pittsburgh, pushed this idea further by using multi-modal datasets that not only include images, but also include translations in a high-resource language. This is an interesting extension as parallel data between speech from an unwritten language and translations of that speech signal in another language can easily be collected [14].

This paper summarizes the accomplishments of the multidisciplinary "Speaking Rosetta" workshop which explored the computational and scientific issues surrounding the discovery of linguistic units (subwords and words) in a language without orthography, through replacing the orthographic transcriptions typically used for training an ASR system by images and/or translations in a well-resourced language. The focus of the project was on discovering intermediate symbolic units and investigating their role in building SLT systems. We concentrated on 4 tasks: two with symbolic units (unit discovery and speech synthesis) and two end-to-end tasks without the need for explicit symbolic units (speech2image and speech2translation).

*Corresponding author: O.Scharenborg@let.ru.nl

†The work reported here was started at JSALT 2017 in CMU, Pittsburgh, and was supported by JHU and CMU via grants from Google, Microsoft, Amazon, Facebook, Apple. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). OS was partially supported by a Vidi-grant from NWO (276-89-003). PG was funded by the French ANR and the German DFG under grant ANR-14-CE35-0002 (BULB project). MD, EL, RR and ED were funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), and ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*.

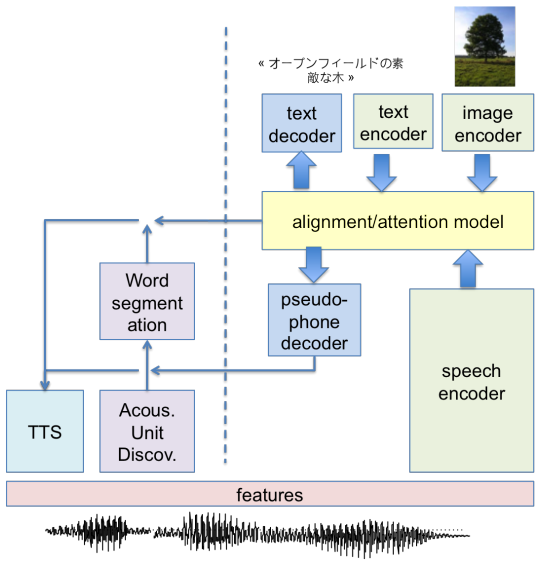

Fig. 1. Functional blocks of the "Speaking Rosetta" project

2. OVERVIEW OF "SPEAKING ROSETTA"

Figure 1 shows a visual representation of the end-to-end systems, and structure, of the Rosetta project. The unit discovery strand (see Section 3.1) focused on discovering 'acoustic units' in the form of articulatory features or (pseudo) phones from raw speech. These acoustic units were used to build speech synthesis systems (Section 3.2), to transform the speech input into symbolic units (pseudo words or pseudo phones) and these units were used for several end-to-end tasks. The end-to-end tasks (see Section 4) used an encoder-decoder framework to translate speech or retrieve images from speech. Alignment/attention models were taken advantage of. Two of these end-to-end tasks are highlighted below: speech to translation and speech to image retrieval.

2.1. Databases

Five multi- and unimodal databases were used. The **Mboshi** (Bantu language spoken in Congo-Brazzavile) corpus[1] consists of 5k speech utterances (approximately 4 hours of speech) in Mboshi aligned to French text. The data set also contains linguists' transcriptions in Mboshi in the form of a non-standard graphemic form close to the language phonology [1, 15].

The **FlickR-real speech** database is a tri-modal (speech, translated text, images) corpus. The FlickR corpus contains 5 different natural language text captions (obtained using Amazon Mechanical Turk; AMT) for each of 8000 images captured from the FlickR photo sharing website. AMT was also used by [16] to obtain 40K spoken versions of the captions. We augmented this corpus by adding Japanese translations (Google MT) for all 40K captions, as well as Japanese tokenization.

SPEECH-COCO-synthetic [17, 18] is an augmentation of MSCOCO [19] which consists of 123,287 images with five different descriptions per image. We generated speech captions using text-to-speech (TTS) synthesis resulting in 616,767 spoken captions (more than 600h) paired with images. Disfluencies and speed perturbation were added to the signal in order to make it sound more natural.

The **How-To dataset** is an English open domain instructional videos (uploaded by users with personal video recorders) dataset of about 480 hours of speech. Each video is broken down into short utterances of about 8-10 seconds each. Transcriptions consist of summaries of what was spoken. The cleanest 45 hours out of the 480 hours were used.

From the **Spoken Dutch Corpus** (CGN, [20]), 64 hours of read speech were used.

2.2. Evaluation

The two types of task, i.e., linguistic unit discovery and end-to-end, were evaluated using a battery of tests which include qualitative measures, e.g., the MCD (see Section 3.2 [21]) and the ABX task (which compares the similarity between discovered units and ground truth labels or between different types of acoustic features; [22, 23]) for the evaluation of the discovered units, and quantitative measures, such as BLEU score, error rates, and word discovery metrics (see for more details [24, 25]).

2.3. XNMT Toolkit

The end-to-end systems used during the project were built using the neural machine translation toolkit XNMT [26], which was greatly improved during the course of this project. XNMT is a sequence-to-sequence neural network toolkit which reads in a sequence of (variable-length) inputs, and then generates a different sequence of (variable-length) output. It consists of a library of standard components. The library is designed so that existing components can be easily re-arranged to run new experiments, and new components can be easily added. Available components are categorized as embedders (e.g., onehot, linear, and continuous vector embedders), encoders (e.g., CNN, LSTM, and pyramidal LSTM encoders), attention models (e.g., dot product, bilinear, and MLP attention models), decoders (e.g., a RNN decoder applied to the state vector of the encoder), and error metrics (e.g., BLEU, cross-entropy, word error rate).

3. TASKS WITH SYMBOLIC UNITS

3.1. Unit discovery

Three different unit discovery systems were implemented that used out-of-domain languages to help unit discovery through (almost) zero-shot adaptation.

In the **unsupervised phoneme discovery - Bayesian acoustic unit discovery (AUD)** approach, pseudo-phones

[1]The dataset will be made available for free by ELRA; its current version is online at: https://github.com/besacier/mboshi-french-parallel-corpus

were generated from the AUD system of [3] with two major modifications. First, the truncated Dirichlet process of [3] was replaced by a symmetric Dirichlet distribution, which provides a good and yet simple approximation of the Dirichlet Process [27]. Second, to cope with larger databases, the Variational Bayes Inference algorithm originally used in [3] was replaced with the faster Stochastic Variational Bayes Inference algorithm. Experiments showed that these modifications, while considerably speeding up the training, yielded negligible drop in accuracy. Also, an extension of this model was explored: the AUD model was embedded into a Variational Auto-Encoder leading to a specific case of the recently developed Structured Variational Auto-Encoder model [28][2].

The **universal articulatory features and phoneme inventory discovery** approach aimed at deriving phone-like units using the setup presented in [29]. It consists of three steps: 1) Detection of pseudo-phone boundaries 2) Extraction of language-universal articulatory features (AFs) for each segment. 3) Clustering of the segments based on the extracted AFs. Seven articulatory feature detectors using different network architectures were trained using data from multiple source languages, and evaluated cross-lingually. Results indicated that the LSTM-based feature extractors showed an improved multilingual performance compared to [29], but they did not perform as good as their feed-forward neural network based counterparts crosslingually. Using k-means, segments were clustered based on the extracted AFs of each segment. Estimating the number of classes k is an open question for future research.

The **cross-language definition of units** approach [30] uses linguistic knowledge of the low-resource language and a semi-supervised training paradigm to build an ASR system for a low-resource language through the adaptation of an ASR system of a high-resource language. Crucially, phones that are present in the low-resource language but not in the high-resource language need to be created. This is done through a linear extrapolation between existing acoustic units in the high-resource ASR system's soft-max layer after which the acoustic units are iteratively retrained using all utterances or only those that have the best score according to four different criteria: ASR score, the MCD score from a TTS system (see Section 3.2), translated text retrieval score, and their combination. The experiments showed that in order to train acoustic units using self-labelled data, training utterances are needed that capture multiple aspects of the speech signal.

3.2. "TTS without T"

Text-to-speech (TTS) technology was used to generate speech from unit sequences, and to evaluate the quality of the discovered unit inventories. Since this project concerns languages without orthography, TTS systems need to be built using discovered units rather than text (dubbed "TTS without T"). The TTS system used is Clustergen [21]. Clustergen works well with small corpora because it treats each frame of the training corpus as a training example, rather than each segment. This makes it suitable for our low-resource scenario. The input to Clustergen is a waveform file plus symbolic sequences of "phones"; the output is a simple synthesizer and a Melcepstral distortion measure (MCD) [31] on held out data. MCD measures the average distance between the log-spectra of the synthetic and natural utterances, and has been demonstrated to be an extremely sensitive measure of the perceived naturalness of speech utterances, e.g., an MCD difference between two synthesis algorithms of 0.3 (on the same test corpus) is usually perceptible by human listeners as a significant difference in perceived naturalness [21].

TTS was used to generate speech in two tasks. The first task is a new speech technology task, which we call **image2speech** [32]. Image2speech is similar to automatic image captioning, but can reach people whose language does not have a natural or easily used written form. The image2speech pipeline consists of a VGG16 visual object recognizer which converts each image into a sequence of feature vectors. XNMT accepts image feature vectors as inputs, and generates speech units as output, which were then sent to the TTS. Four types of intermediate speech units were tested: 1) L1-words and 2) L1-phones (generated using a same-language ASR, which provides an upper bound performance); 3) L2-phones from the cross-language definition of units approach and 4) pseudo-phones generated using AUD (see Section 3.1 for both). Results showed that the image2speech system is able to generate a phone string that is composed entirely of intelligible words, sequenced in an intelligible and semantically reasonable sentence.

In the second task, a proof-of-concept **foreign-text-2-speech** end-to-end system was build using XNMT which translates French words (text) into Mboshi phones which were either (1) true phones (2) or pseudo-phones obtained via AUD (see Section 3.1). These phone sequences were then sent to the TTS system. On a development set of 514 utterances BLEU4 scores at the character level were of 31.95% with true phones and 8.32% with pseudo-phones[3].

3.3. Speech and image to text (and summarization)

The speech-and-image2text system uses multi-modal information consisting of speech and videos to improve standard (supervised) ASR (this approach is thus also useful for high-resource languages). From the videos, object and scene features are extracted and used to adapt a sequence-to-sequence model (using the Pyramidal encoder by [33]) to the visual features. Results showed that adding the visual features helps the model convergence and guides the training in the earlier epochs, compared to an HMM-DNN model. The sequence-to-sequence model is able to jointly learn the audio visual

[2]The source code of both AUD models is available via https://github.com/amdtkdev/amdtk

[3]TTS speech samples are available via https://github.com/JSALT-Rosetta/Illustrations/blob/master/TTS/mboshi/

System	Prec	Recall	F
Segmental DTW Baseline [37]	31.9	13.8	19.3
Attention (fr-mb)	36.5	46.1	40.7
Attention (mb-fr)	36.3	46.6	40.8

Table 1. Speech-to-translation: Word boundary detection results (Mboshi5k corpus) from pseudo phones

features, the acoustic and language models, requires no extra preprocessing for noisy data, does not require precomputed alignments, and is efficient even with long utterances.

4. END-TO-END TASKS

4.1. Speech-to-translation

End-to-End speech translation, i.e., translation from raw speech without any intermediate transcription [34, 35], is attractive for language documentation, which often uses corpora made of audio recordings aligned with their translation in another language (no transcript in the source language) [1, 14]. Here, XNMT was used to build end-to-end speech translations systems on FlickR (English-to-Japanese) and Mboshi-to-French. The obtained BLEU4 scores at the character level were 30.99% and 22.36% on the development sets of FlickR and Mboshi, respectively. Although these results are rather low for a pure translation task, these systems show that end-to-end models are able to encode some regularities in the speech signal in order to decode predictable sequences of characters in a target language.

Secondly, an attention-based Neural Machine Translation (NMT) model [36] was trained between phones in Mboshi and text in French, while soft-alignment probability matrices generated by the attention mechanism, were extracted. These alignments were post-processed to segment a sequence of symbols in Mboshi into words. While [36] applied their method to true phones (gold phonemes transcribed by linguists), here segmentation through attention from a pseudo-phone sequence obtained using AUD (see Section 3.1) was investigated. Table 1 shows that the word boundary detection results of the attention-based system outperformed those of a pure speech-based baseline which used pair-matching using locally sensitive hashing applied to PLP features and then grouped pairs using graph clustering [37]. Moreover, a reverse model (French-Mboshi) slightly improved word segmentation compared to (Mboshi-French). Implementation of a bilingual loss is probably an interesting future work.

4.2. Speech-to-Image

Speech-to-image is a relatively new task [11, 12, 13]. A speech-to-image system learns to map images and speech to the same embedding space, and retrieves an image using spoken captions. While doing so, it uses multi-modal input to discover speech units in an unsupervised manner, similar to how children acquire their first language. Our speech-to-image system (based on the implementation of [13]) was implemented using XNMT. Four types of acoustic features were compared: Mel-frequency Filterbanks (baseline, similar to [16] but with added speaker-dependent mean-variance normalization on the features before zero-padding/truncation), the pseudo-phones generated by the AUD system [3] (which were downsampled by a factor of 9 along the phone dimension to fit the input of the DNN), Multilingual Bottleneck features (MBN), and Cochleagram Features generated by the Resonant Tectorial Model developed by [38].

Feature type	R@1	R@5	R@10
Mel-filterbank	0.0096	0.047	0.0856
Multiling. Bottleneck	**0.013**	**0.053**	**0.0994**
AUD (one epoch)	0.0012	0.0044	0.0112
Cochleagram	0.0008	0.005	0.0104

Table 2. Speech-to-image retrieval results (Recall@N) for the tested input speech features

Table 2 shows the results for the four features evaluated with Recall@N. The MBN feature is superior to all other acoustic features, and shows over 1 percent improvement on the Filterbank baseline for the recall@10 score.

5. CONCLUDING REMARKS

The "Speaking Rosetta" JSALT 2017 project laid the foundation for a new research area "Unsupervised multi-modal language acquisition". It showed that it is possible to build useful SLT systems without any textual resources in the language for which the SLT is built, in a way that is similar to that of how infants learn a language. 1) The "Speaking Rosetta" project showed that zero-shot adaptation, i.e., unsupervised learning of speech units, is possible, and can be improved by using information extracted from well-resourced languages. The discovered units are meaningful as shown by their usefulness in upstream tasks such as word discovery, image retrieval, and speech translation tasks. We have presented the first attempt to discover spoken term from speech using an attention matrix; the performance of this approach is better than all the baselines evaluated in the same conditions. 2) TTS has proven to be a useful tool in the evaluation of discovered units of different types, and can be used to evaluate how well a particular set of units correlates with acoustic features. "Units" we have tested include articulatory features, AUDs, and cross-language adapted phones. 3) The unit-discovery and the end-to-end systems were successfully combined into several working proof-of-concept end-to-end demos. 4) We showed that audio and video information can be fused to improve speech summarization without going through text. 5) Finally, a pipeline of metrics as well as dedicated datasets were created to fuel reproducible researches in this new emerging domain.

6. REFERENCES

[1] G. Adda et al., "Breaking the unwritten kanguage barrier: The Bulb project," in *Proceedings of SLTU*, Yogyakarta, Indonesia, 2016.

[2] A. Jansen et al., "A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of early language acquisition," in *Proceedings of ICASSP*, 2013.

[3] L. Ondel, L. Burget, and J. Černocký, "Variational inference for acoustic unit discovery," in *Procedia Computer Science*, 2016, pp. 80–86.

[4] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of ACL on Human Language Technologies: Short Papers*, 2008, pp. 165–168.

[5] A. S. Park and J. R. Glass, "Unsupervised Pattern Discovery in Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[6] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proceeding of ICASSP*, 2010, pp. 4366–4369.

[7] A. W. Tanja Schultz, "Experiments on cross-language acoustic modelling," in *Proceedings of Interspeech*, 2001.

[8] J. Lööf, C. Gollan, and H. Ney, "Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a polish speech recognition system," in *Proceedings of Interspeech*, 2009.

[9] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proceedings of SLT*, 2012.

[10] H. Xu, V. Do, X. Xiao, and E. Chng, "A comparative study of bnf and dnn multilingual training on cross-lingual low-resource speech recognition," in *Proceedings of Interspeech*, 2015, pp. 2132–2136.

[11] D. Harwarth and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proceedings ASRU*, 2015, pp. 237–244.

[12] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of ASRU*, 2017.

[13] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing System*, 2016, pp. 1858–1866.

[14] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and A. Rialland, "Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app," in *Proceedings of SLTU*, Yogyakarta, Indonesia, May 2016.

[15] P. Godard et al., "A very low resource language speech corpus for computational language documentation experiments," in *arXiv:1710.03501*, 2017.

[16] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proceedings of ASRU*, Scottsdale, Arizona, USA, 2015, pp. 237–244.

[17] L. Besacier, "Speech-coco," https://persyval-platform.univ-grenoble-alpes.fr/DS80/detaildataset.

[18] W. Havard, L. Besacier, and O. Rosec, "Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set," in *International Workshop on Grounding Language Understanding (GLU), Satellite of Interspeech 2017*, 2017.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, Zürich, 2014, Oral.

[20] N. Oostdijk, W. Goedertier, F. V. Eynde, L. Boves, J.-P. Martens, M. Moortgat, and H. Baayen, "Experiences from the spoken dutch corpus project," in *Proceedings of LREC, Las Palmas de Gran Canaria*, 2002, pp. 340–347.

[21] A. W. Black, "CLUSTERGEN: A statistical parametric speech synthesizer using trajectory modeling," in *Proceedings of ICSLP*, 2006, pp. 1762–1765.

[22] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the Minimal-Pair ABX task (I): Analysis of the classical MFC/PLP pipeline," in *Proceedings of Interspeech*, 2013.

[23] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise," in *Proceedings of Interspeech*, 2014.

[24] B. Ludusan, M. Versteegh, A. Jansen, G. Gravier, X.-N. Cao, M. Johnson, and E. Dupoux, "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," in *Proceedings of LREC*, 2014.

[25] E. Dunbar, X. Nga Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *Proceedings of ASRU*, 2017.

[26] G. Neubig, "Xnmt," https://github.com/neulab/xnmt/.

[27] K. Kurihara, M. Welling, and Y. W. Teh, "Collapsed variational Dirichlet process mixture models," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007, vol. 20.

[28] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams, "Composing graphical models with neural networks for structured representations and fast inference," in *Neural Information Processing Systems*, 2016.

[29] M. Müller, J. Franke, S. Stüker, and A. Waibel, "Improving phoneme set discovery for documenting unwritten languages," *Elektronische Sprachsignalverarbeitung (ESSV) 2017*, 2017.

[30] O. Scharenborg, F. Ciannella, S. Palaskar, A. Black, F. Metze, L. Ondel, and M. Hasegawa-Johnson, "Building an asr system for a low-resource language through the adaptation of a high-resource language asr system: Preliminary results," in *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[31] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *Proceedings of SSW5, Pittsburgh, PA*, 2004, pp. 31–36.

[32] M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, "Image2speech: Automatically generating audio descriptions of images," in *Proceedings of ICNLSSP, Casablanca, Morocco*, 2017.

[33] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[34] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *NIPS workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[35] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech," *arXiv preprint arXiv:1703.08581*, 2017.

[36] M. Zanon Boito, A. Berard, A. Villavicencio, and L. Besacier, "Unwritten languages demand attention too! word discovery with encoder-decoder models," in *Proceedings of ASRU*, 2017.

[37] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proceedings of ASRU*, 2011, pp. 401–406.

[38] J. Allen and D. Sen, "Is tectorial membrane filtering required to explain two tone suppression and the upward spread of masking?," in *Mechanics of Hearing*, 1999, p. 451.