post-processing speech-to-text transcripts

STATE-OF-THE-ART SURVEY

Steven Walker | ZHAW InIT | Jan 2019

Abstract

Automatic Speech Recognition (ASR) has been a major research topic for decades and many different approaches exist. This survey in transcription post-processing (i.e. all processes apart from translating phonemes to words, depending on the system), aims to give a high-level insight into the state-of-the-art approaches. First the basic ASR system is described with emphasis on the decoder part. This is necessary, not only for a basic comprehension of ASR systems, but also since some of the methods in the following chapters are included in the basic structure (e.g. language identification is performed in the language model) and aren't individual external components.

In a second step the different methods (chapters 2.3 to 2.6) that can be considered tasks of post processing transcripted speech are explained, including their state-of-the-art methods and a short insight into the development. In 2.7 conversational analysis is briefly mentioned, since some applications want transcriptions to describe a conversation as precise as possible.

Overall the emergence of Deep neural networks (DNNs) have led to significant improvements in all processes, as well as to a decrease in complexity, since often hand-engineered tasks have become obsolete.
Furthermore, new DNN models like Attention based ones (based on human attention mechanisms) introduced in 2015 are showing promising results and a strive to simpler End-to-End DNN models is a goal in all ASR research areas. Benefits are a decrease in complexity and more possible applications, e.g. multi language/dialect handling as mentioned in chapter 2.5 Language identification (LID).

Table of content

1. Introduction

Enabling computers to recognize human speech has been a central problem of artificial intelligence for decades. Numerous applications emerged since ASR performance has reached a practical level, ASR is used in health care (e.g. medical documentation), military (e.g. cockpit control in planes and helicopters), telephony (e.g. voice search, text dictation, but also call transcription etc.) and more recently for applications like providing subtitles for video (e.g. YouTube), or customer support applications using conversational-agents/chatbots.

The mentioned improvements, enabling many applications from above, are mostly due to the emergence of deep artificial neural networks (DNNs), more specific, due to the introduction of improved Recurrent Neural Networks (RNNs) in the late 2000s. -These models also brought improvements to domains like machine translation and handwriting recognition, just to name a few. In the last decade, new DNN models have emerged, like attention based neural networks and complexity reducing End-to-End models are further improving ASR performance.

State of the Art ASR systems achieve word error rates (WER) as low as 5.8%, but there is still work to be done. The training sets often don't resemble real live data quality, data sets are often too clean, don't have more than one speaker, or don't contain multi-lingual words just to name a few. (Su, 2018)

To make an ASR model more generic and robust, further research in speaker diarization (chapter 2.4), language identification (chapter 2.5) and vocabulary enhancement (chapter 2.6) is required.

With improving ASR systems, more requirements are set, e.g. punctuation restoration (chapter 2.3), speaker adaption/characterization, recognizing overlapped speech as in e.g. meetings and robustness to dialect variation, just to name a few.

Furthermore, less popular languages haven't yet had the attention from researchers, that e.g. English or mandarin had and still require research on language specific features/syntax e.g. the compound words seen in the German languages (see chapter 4.1).

Often in this survey technologies and model examples from Google are showed or explained, this isn't due to preference, but due to extensive and up-to-date information provided through Googles blog posts. The use of blogs is has become more common (e.g. Google, Baidu and more), especially in this time, where every year new findings are reported, and companies are constantly improving their products. Of course, these blogs do not show the same depth as a paper and often some techniques are probably kept secret for competitive reasons, however they still offer insight in the direction the industry is going, problems encountered on their way and technologies chosen for the product or service. Blogs are also listed as sources but aren't immediately distinguishable from other sources.

1.1 OBJECTIV OF THIS PAPER

This survey on ASR post-processing techniques is also an introduction into post-processing ASR transcriptions and aims at giving researchers that are new in the field some insight into state-of-the-art methods. Therefore, at the end of each chapter, Sources and relevant papers are listed, including the year and if existing the conference at which it was presented, providing a quick relevantness assessment option.

1.2 ADVISABLE FUNDAMENTAL KNOWLEDGE

The field of computational linguistics is wide, deep and extensive, since up until the 21. Century other Algorithms and models were used, compared to the Deep neural networks used today.

The internet contains extensive information on all topics. They cover many topics in speech recognition. Following is a list of advisable fundamental knowledge for people who wish to work with ASR.

- NIST Rich Transcription evaluations
- Transcription guidelines
- Transliteration
- Orthographic transcription and Phonetic transcription
- Phonetics (e.g. intonation)
- International Phonetic Alphabet (IPA)
- Morphology, morphemes
- Feature Extraction, acoustic model, language model
- N-Grams
- Part of Speech Tagging (POS)
- Word Embedding (Word2vec, GloVe)
- Statistical parsing
- Lexical Semantics
- Speech Synthesis
- Hidden Markov and Maximum Entropy Models
- WordNet, WikiText and Switchboard Datasets etc.
- Basic features in speech processing (prosodic features):
  pitch, duration, intensity, voice quality, signal to noise ratio, voice activity detection, user switch detection and strength of Lombard effect
- Basic Computational linguistics

- CNNs, LSTMs, RNNs
- Distilling Knowledge in Neural Network ensembles
- Attention in neural networks
- Conversational Analysis (chapter 2.7)
- MFCC, i-vector, d-vector

2 State of the art

The field of Automatic Speech Recognition (ASR) systems has experienced a paradigm shift in the last 10 to 20 years and still every year new models and concepts appear, proposing new neural networks and model structures. This introduction into the ASR decoder gives an overview over recent developments.

Hidden-Markov Models (HMM) and Gaussian Mixed Models (GMM) were used for the first reasonably good ASR systems. Google Voice transcription (launched 2009), used a GMM as an acoustic model (see ASR Decoder) until they opted for LSTM RNNs in 2012, when Long Short Term Memory (LSTM) RNNs revolutionized the field of speech recognition (Beaufays, 2015).

Conventional ASR systems use multiple different artificial neural networks or HMMs/GMMs depending on the purpose of the system. Most applications for ASR are structured as *hybrids*, using individual networks for each component. Conventional systems achieve word error rates (WER) of around 16%.

In the traditional approach, the speech recognizer consists of a *feature extractor* (encoder) and a decoder, which again consists of an *acoustic model*, a *pronunciation model* or *dictionary* and a *language model.* In general, each of these modules could be built separately and then integrated together to perform ASR.
Google defines the ASR system as having three components, an acoustic model, a pronunciation model and a language model, all of which are independently trained. (Pundak and Sainath, 2016)

The *feature extractor* cuts the audio signal into frames (e.g. 10ms) and applies the mel-frequency cepstrum, generating feature vectors (MFCCs). The *decoder* then stepwise transforms the feature vectors to phonemes, then to words, and finally to sentences, using acoustic, pronunciation and language model (see Figure 1). The *decoder* calculates the most likely sequence of words for each utterance, given the probabilities provided by the acoustic and language models.

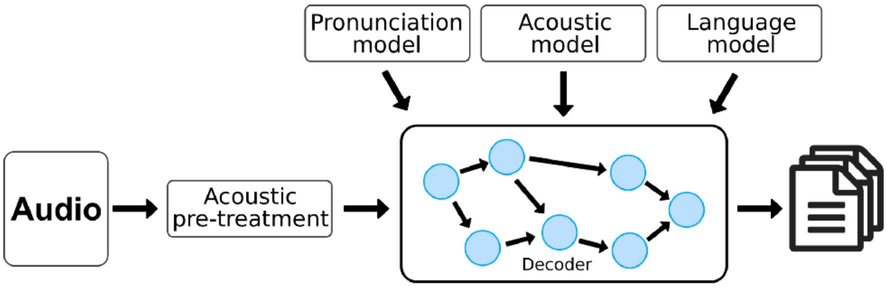


Figure 1, A classic modern ASR model (Linagora, 2018)

This survey on post-processing transcription focusses on processes applied on the output of an ASR model. However, in many cases it is hard to draw a clear line.

2.1 ASR DECODER

As mentioned above, the decoder consists of an acoustic model, a pronunciation model and a language model.

The acoustic model is the art of transcribing the MFCC feature vectors from the feature extractor to phonetic language. The modern approach is to transform MFCCs using long-short-term-memory (LSTM) networks, predicting the string of phonemes most likely. Newer approaches take convolutional neural networks (CNNs) to directly analyze the spectral image, using their strength in 2-D pattern recognition. CNNs are effective models for reducing spectral variations and modeling spectral correlations in acoustic features. With this approach, feature extraction is incorporated into the acoustic model.

These LSTM RNN have gained a boost of popularity since 2014 and lead to great improvements in sequence to sequence applications. However, new Attention based systems are on the verge of replacing simple RNNs and LSTMs and should be considered for new models: “RNN have the days counted in all applications, because they require more resources to train and run than attention-based models” (Eugenio, 2018).

The trained *acoustic model* outputs the phonemes most likely, according to the feature vectors.
The *pronunciation model* provides mapping relationships between the acoustic modeling units and language modeling units. Different pronunciations of words, make the mapping a challenge.

The *language model* is a key component in many NLP applications, especially ones that generate text as an output. While older models used grammars to specify what word sequences are allowed, modern models are statistical. Statistical language models are recommended for free-form input where the user could say anything in a natural language. They only require a database with possible sentences, and thus less engineering effort than grammars. For example, if numbers like “twenty one” and “thirty three” are listed, a statistical language model will allow “thirty one” with a certain probability as well.
In its essence a statistical language model is an encoding of statistical information about co-occurrence of words or word patterns. Or as written on Wikipedia: “A statistical language model is a probability distribution over sequences of words. Given such a sequence, say of length m, it assigns a probability $P(w\_1, \ldots, w\_m)$ to the whole sequence. [...] The language model provides context to distinguish between words and phrases that sound similar. For example, in American English, the phrases "recognize speech" and "wreck a nice beach" are pronounced almost the same but mean very different things.” (‘Language model’, no date).

Traditional statistical language models use n-gams, which calculate the probability of each word, given the n previous words. E.g. for the sentence “The train arrived late” and n=3: $P(The, train, arrived, late) \approx P(The \mid <s>, <s>)\, P(train \mid The, <s>) \ldots$

$$\ldots P(arrived \mid train, The)\, P(late \mid arrived, train)\ \ P(</s> \mid late, arrived)$$

Start- and End-Of-Sentence markers are typically denoted as <s>, </s> or in part of speech tagging (EOS).

N-grams are very common and many ASR systems in production are still based on N-grams, in some cases enriched with an additional RNN at the output (Jyothi, 2017). N-grams get trained with large corpora of text and compute normalized counts to get word probabilities. To enable n-grams to handle unseen word contexts, smoothing models are applied.

Modern well-established approaches use LSTM RNNs or bidirectional RNNs in connection with a Viterbi search algorithm. RNNs have the ability to model more context than the n-gram models by taking much more previous occurred words into account. Bidirectional networks can also take context from following words into account. Finally, attention-based networks are able to focus on specific context of the word being predicted, making them more efficient and faster trainable. “BLSTM increases performance by 15%-20% compared to the previous generation of DNN-based modeling methods.” (Alibaba Cloud iDST, 2016)

Papers on acoustic models:

Maha Elbayad, Laurent Besacier, Jakob Verbeek, “Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction” CoNLL 2018

Has¸im Sak, Andrew Senior, Franc¸oise Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling” INTERSPEECH 2014

Video on pronunciation model:

Preethi Jyothi from Microsoft, speaks at the IIT Bombay “Automatic Speech Recognition – An Overview” https://www.microsoft.com/en-us/research/video/automatic-speech-recognition-overview/

Papers on language models:

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, Sanjeev Khudanpur, “Recurrent neural network based language model” INTERSPEECH 2010

Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, Stanley Chen, “Bidirectional recurrent neural network language models for automatic speech recognition” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney “LSTM Neural Networks for Language Modeling” INTERSPEECH 2012

George Saon, Hong-Kwang J. Kuo, Steven Rennie and Michael Picheny, “The IBM 2015 English Conversational Telephone Speech Recognition System” INTERSPEECH 2015

Journal article on https://machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/ Jason Brownlee, "Gentle Introduction to Statistical Language Modeling and Neural Language Models" posted on the 1. of November 2017.

Language model tools (located at the end of the article):

https://en.wikipedia.org/wiki/Language_model#cite_ref-bengio_6-1

Hidden Markov Model Toolkit:

http://htk.eng.cam.ac.uk/

2.2 END-TO-END ASR SYSTEMS

End-to-End systems are the latest development and show promising results, while decreasing complexity of the whole ASR model. These systems incorporate all components into one neural network model (e.g. they can directly transform the acoustic signal to word sequences), such as the Listen-Attend-Spell (LAS) architecture. End-to-end acoustic-to-word speech recognition models have recently gained popularity because they are easy to train, scale well to large amounts of training data, and don't require a lexicon.
"Training independent components creates added complexities and is suboptimal compared to training all components jointly" (Sainath and Wu, 2017).
Google has achieved a good WER of 5.6% with a LAS End-to-End architecture introduced in December 2017 (see Figure 2). Additionally, they report it to be smaller than conventional models, since it contains no separate language and phonetic model. Furthermore, google sees a lot of potential also in Multilanguage recognition, since the system isn't based on a dictionary. However, they also report that currently, these models cannot yet process speech in real time.

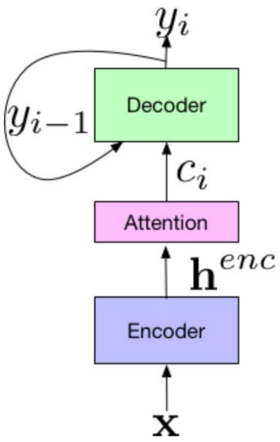


Figure 2, Components of the LAS End-to-End Model. (Sainath and Wu, 2017)

Baidu researchers are eagerly developing End-to-End models as well and released Deep Speech 2 in December 2015 (Amodei *et al.*, 2015). This model is interesting, since

Amodei et al. claim, that it works equally well with English as it works with Mandarin. "Because it replaces entire pipelines of hand-engineered components with neural networks, end-to-end learning allows us to handle a diverse variety of speech including noisy environments, accents and different languages" (Amodei *et al.*, 2015)

End-to-End ASR systems:

Kartik Audhkhasi et al. "Building Competitive Direct Acoustics-to-Word Models for English Conversational Speech Recognition" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018

Ying Zhang et al. 2017 "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks" 2017

W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell" 2015

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper et. al. "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin" 2015

Eric Battenberg, Jitong Chen, Rewon Child, Adam Coates, Yashesh Gaur, Yi Li, Hairong Liu, Sanjeev Satheesh, David Seetapun, Anuroop Sriram, Zhenyao Zhu, "Exploring Neural Transducers for End-to-End Speech Recognition" IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2017

C.C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R.J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski and M. Bacchiani, "State-of-the-art Speech Recognition With Sequence-to-Sequence Models" ICASSP 2018

2.3 RESTORING PUNCTUATION

Most output of ASR systems don't contain punctuation, although punctuation often plays a crucial role in understanding the semantics of a sentence. Restoring punctuation was heavily researched in the past few years and still is. Usually the language model contains the model responsible for restoring punctuation, but it is also possible to place the model at the output of a ASR system.
Earlier approaches made use of predefined grammar rules, others used decision trees and hidden event N-grams, while some simply allow the dictation of punctuation symbols. But significant accuracy wasn't achieved until LSTMs were introduced. Many papers suggest different approaches to the task (see papers listed below). While some approaches only use lexical features, some additionally take acoustical features into account as well. Tal Levy et al. analyzed the effect of pitch, intensity and pause duration in punctuation detection in 2012 (not considering context at all): "Results show that 87% of full-stops were detected, with only 14% false alarms. Nevertheless, since most commas are realized with no pitch breaks, only 54% of the commas were detected, with 35% false alarms." (Levy, Silber-Varod and Moyal, 2012)

However, some researchers report inaccuracies when the user makes pauses in unnatural moments.

Newer approaches make use of LSTMs. LSTMs trained with large text corpuses also brought improvement to this application. Popular approaches use hybrid models taking lexical and acoustical information, such as pitch, intensity, pause, energy and speaker switch information, into account. Three architectures are common (Figure 3).

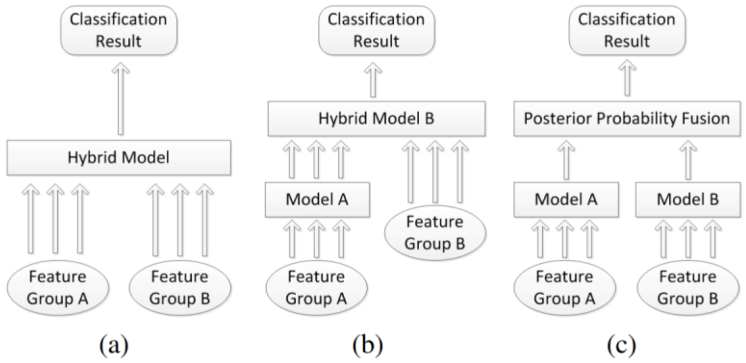
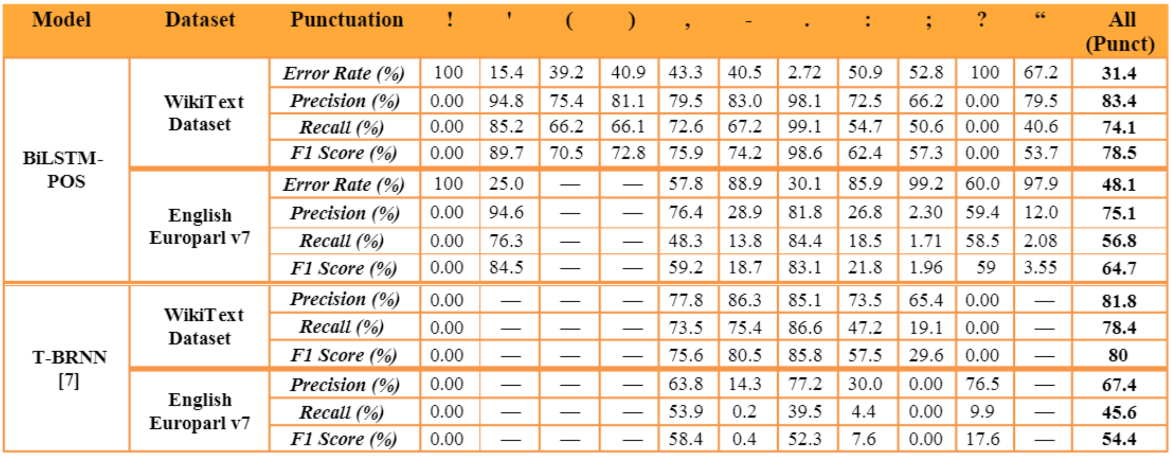


Figure 3, Three types of model structures. Feature Group A and B are placeholders for a lexical and an acoustical model. (Che *et al.*, 2016)

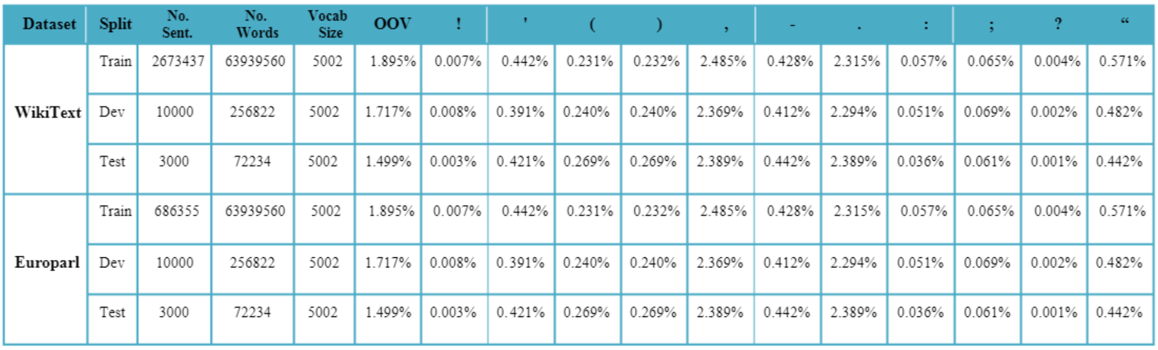
Google's voice transcription went through this transition: "The old system relied on hand-crafted rules or "grammars," which, by design, can't easily take textual context into account. For example, in an early test our algorithms transcribed the audio "I got the message you left me" as "I got the message. You left me." To try and tackle this, we again tapped into neural networks, teaching an LSTM to insert punctuation at the right spots." (Beaufays, 2015)

Chin Char Juin et al. presented their bidirectional LSTM network with part-of-speech (PoS) tagging in November 2017, including a comparison to the LSTM network of (Tilk and Alumäe, 2016). A set of 11 different punctuation marks were predicted, which is larger than most sets previously reported in the literature. However, exclamation and question marks performed very badly, according to the authors, the lack of training data containing those marks could be held accountable. (E.g Figure 5 shows an exclamation mark occurrence of 0.007% in the training data (no. of sentences: 2'673'437).
Part of speech tagging was beneficial, especially when identifying apostrophes, since those can usually be determined by the "_POS" part-of-speech tag, which refers to possessives such as 's or 'd.
The model with the WikiText dataset (WikiText is a collection of over 100 million tokens extracted from a set of high quality articles on Wikipedia) achieved a good WER of 31.4% when counting only the punctuations, and an F1-score of 78.5, which is on par with past systems despite having to predict more types of punctuations.



Model	Dataset	Punctuation	!	'	(	)	,	-	.	:	;	?	“	All (Punct)
BiLSTM-POS	WikiText Dataset	*Error Rate (%)*	100	15.4	39.2	40.9	43.3	40.5	2.72	50.9	52.8	100	67.2	**31.4**
		Precision (%)	0.00	94.8	75.4	81.1	79.5	83.0	98.1	72.5	66.2	0.00	79.5	**83.4**
		Recall (%)	0.00	85.2	66.2	66.1	72.6	67.2	99.1	54.7	50.6	0.00	40.6	**74.1**
		F1 Score (%)	0.00	89.7	70.5	72.8	75.9	74.2	98.6	62.4	57.3	0.00	53.7	**78.5**
	English Europarl v7	*Error Rate (%)*	100	25.0	—	—	57.8	88.9	30.1	85.9	99.2	60.0	97.9	**48.1**
		Precision (%)	0.00	94.6	—	—	76.4	28.9	81.8	26.8	2.30	59.4	12.0	**75.1**
		Recall (%)	0.00	76.3	—	—	48.3	13.8	84.4	18.5	1.71	58.5	2.08	**56.8**
		F1 Score (%)	0.00	84.5	—	—	59.2	18.7	83.1	21.8	1.96	59	3.55	**64.7**
T-BRNN [7]	WikiText Dataset	*Precision (%)*	0.00	—	—	—	77.8	86.3	85.1	73.5	65.4	0.00	—	**81.8**
		Recall (%)	0.00	—	—	—	73.5	75.4	86.6	47.2	19.1	0.00	—	**78.4**
		F1 Score (%)	0.00	—	—	—	75.6	80.5	85.8	57.5	29.6	0.00	—	**80**
	English Europarl v7	*Precision (%)*	0.00	—	—	—	63.8	14.3	77.2	30.0	0.00	76.5	—	**67.4**
		Recall (%)	0.00	—	—	—	53.9	0.2	39.5	4.4	0.00	9.9	—	**45.6**
		F1 Score (%)	0.00	—	—	—	58.4	0.4	52.3	7.6	0.00	17.6	—	**54.4**

Figure 4, Per punctuation WER and F1 scores. (Juin *et al.*, 2017)



Dataset	Split	No. Sent.	No. Words	Vocab Size	OOV	!	'	(	)	,	-	.	:	;	?	“
WikiText	Train	2673437	63939560	5002	1.895%	0.007%	0.442%	0.231%	0.232%	2.485%	0.428%	2.315%	0.057%	0.065%	0.004%	0.571%
	Dev	10000	256822	5002	1.717%	0.008%	0.391%	0.240%	0.240%	2.369%	0.412%	2.294%	0.051%	0.069%	0.002%	0.482%
	Test	3000	72234	5002	1.499%	0.003%	0.421%	0.269%	0.269%	2.389%	0.442%	2.389%	0.036%	0.061%	0.001%	0.442%
Europarl	Train	686355	63939560	5002	1.895%	0.007%	0.442%	0.231%	0.232%	2.485%	0.428%	2.315%	0.057%	0.065%	0.004%	0.571%
	Dev	10000	256822	5002	1.717%	0.008%	0.391%	0.240%	0.240%	2.369%	0.412%	2.294%	0.051%	0.069%	0.002%	0.482%
	Test	3000	72234	5002	1.499%	0.003%	0.421%	0.269%	0.269%	2.389%	0.442%	2.389%	0.036%	0.061%	0.001%	0.442%

Figure 5, Statistics of the Databases used. (Out of vocabulary (OOV) shows the amount of words not included in the vocabulary) (Juin *et al.*, 2017)

In 2018 (F. Wang *et al.*, 2018) introduced a Network, based solely on Self-Attention modules, Wang claims it to be superior to traditional methods in simple punctuation and joint punctuation tasks. However, it is to early to see if Attention based models will become state-of-the-art in this domain.

April 2018, Google announced automatic punctuation as a beta function for Google's Speech-to-Text service. The service can transcribe audio in real-time and for English language includes punctuation. Unfortunately, exact information on Google's neural network aren't available, at this time.

Papers on restoring Punctuation:

O. Tilk and T. Alumäe, "LSTM for punctuation restoration in speech transcripts," Interspeech 2015

O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration" Interspeech 2016.

N. Ueffing, M. Bisani, and P. Vozila, "Improved models for automatic punctuation prediction for spoken and written text," Interspeech 2013

D. Zhang, S. Wu, N. Yang, and M. Li, "Punctuation prediction with transition-based parsing." ACL 2013

X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," International Conference on Language Resources and Evaluation (LREC), 2016

Feng Wang, Wei Chen, Zhen Yang, Bo Xu, «Self-Attention Based Network for Punctuation Restoration" International Conference on Pattern Recognition (ICPR) 2018

Alp Oktem, Mireia Farrús, Leo Wanner, "Attentional Parallel RNNs for Generating Punctuation in Transcribed Speech" International Conference, SLSP 2017

Chin Char Juin, Richard Xiong Jun Wei, Luis Fernando D'Haro, Rafael E. Banchs "Punctuation prediction using a bidirectional recurrent neural network with part-of-speech tagging" TENCON 2017

Jiangyan Yi, Jianhua Tao, Zhengqi Wen1, Ya Li1, "Distilling Knowledge from an Ensemble of Models for Punctuation Prediction" INTERSPEECH 2017

Xiaoyin Che, Sheng Luo, Haojin Yang, Christoph Meinel, "Sentence Boundary Detection Based on Parallel Lexical and Acoustic Models" INTERSPEECH 2016

2.4 SPEAKER DIARIZATION

Speaker diarization is the task of determining "who spoke when?" in an audio or video recording that contains an unknown number of speakers. Speaker recognition tasks can be split into text-dependent tasks, typical in password-based access control systems, and into text-independent tasks, like in ASR. As seen in the chapter restoration of punctuation, speaker information can be used to improve transcription quality, e.g. by helping the language model to build speaker dependent sentences and not mix up utterances.

No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary etc.

In the late 1990s, Gaussian mixture models (GMMs) became the dominant approach for text-independent speaker recognition.

Since 1996 the National Institute of Standards and Technology (NIST) periodically organizes Speaker Recognition Evaluations (e.g. SRE16) pushing research in the field of speaker recognition.

Since 2011 factor analysis is used to define a low-dimensional space that models speaker variability. The corresponding vectors are called i-vectors. An i-vector is a fixed-size representation (typically from 400 to 600 dimensions) of a whole utterance. The i-vector approach can be seen as state-of-the-art in the speaker recognition field.

A classic i-vector pipeline can be split into different steps (simplified):

UBM modelling; A GMM or DNN is trained from feature vectors (such as MFCCs or in some cases bottleneck features) and training data consisting of many utterances by different Speakers. The resulting model is known as Universal Background Model (UBM) and is defined by its mean vector (known as supervector) and its covariance matrix.
Statistics computation; Given a trained UBM, the next step is to compute the Baum-Welch statistics for a given utterance.
Total variability subspace training and i-vector extraction; The idea of the Total Variability approach is to project the UBM-supervector from a given utterance into a subspace T, in which the variability of a training dataset is represented. The T projection matrix is trained with a dataset which includes speaker variability for the target task (e.g. speaker diarization, but also speaker recognition, identification or clustering etc.). In a next step the i-vectors can be extracted, containing speaker information on the utterance they represent.
Classification; Finally, classification is performed on the computed i-vectors. Unseen data points are classified as the speaker closest in the i-vector space. Possible classification algorithms include cosine distance scoring, Linear Discriminant Analysis (LDA), Probabilistic LDA (PLDA), Bayesian information criterion (BIC) or cross-likelihood ratio (CLR).

Nevertheless, using MFCC feature vectors for speaker characterization has been criticized, due to the them not exploiting any specific voice-related characteristics of the speech signal. Some features (as e.g. pitch information) are even knowingly neglected.

Recognizing the downsides to MFCCs, Yanick Lukic et al. (ZHAW) proposed a CNN model that learns to extract speaker identifying information from spectral images, without the need for handcrafted features, while performing equally well (Lukic *et al.*, 2016).

With the rise of LSTMs, neural network-based audio embeddings, also known as d-vectors, have demonstrated superior speaker verification performance. By combining LSTM-based d-vector audio embeddings with recent work in non-parametric clustering, state-of-the-art speaker diarization system are obtained. “Our system is evaluated on three standard public datasets, suggesting that d-vector based diarization systems offer significant advantages over traditional i-vector based systems. We achieved a 12.0% diarization error rate on NIST SRE 2000 CALLHOME, while our model is trained with out-of-domain data from voice search logs." (Q. Wang *et al.*, 2018)

Google researchers proposed fully supervised speaker diarization approach in October 2018, named unbounded interleaved-state recurrent neural networks (UIS-RNN) and managed to outperform their previous work (section above). The Term "fully" implies, that all components, including number of speakers, are trained in supervised ways, so that they can benefit from increasing the amount of labeled data available. (Wang, 2018)
They also use d-vectors and RNNs, as well as a *distance-dependent Chinese restaurant process* (ddCRP) to accommodate an unknown number of speakers. The results are promising: They achieved a 7.6% diarization error rate on NIST SRE 2000 CALLHOME, which is better than the state-of-the-art method using spectral clustering. Moreover, they're method decodes in an online fashion while most state-of-the-art systems rely on offline clustering.
Although the system performs well, they see two more improvements for further research: "First, we are refining our model, so it can easily integrate contextual information to perform offline decoding. This will likely further reduce the DER, which is more useful for latency-insensitive applications. Second, we would like to model acoustic features directly instead of using d-vectors. In this way, the entire speaker diarization system can be trained in an end-to-end way." (Wang, 2018)

Open source speaker diarisation software:
https://en.wikipedia.org/wiki/Speaker_diarisation

Papers on speaker diarisation/clustering:

Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P. "Front-End Factor Analysis for Speaker Verification" IEEE Transactions on Audio, Speech, and Language Processing. 2011

Xavier Anguera ; Simon Bozonnet ; Nicholas Evans ; Corinne Fredouille ; Gerald Friedland ; Oriol Vinyals, "Speaker Diarization: A Review of Recent Research" IEEE Transactions on Audio, Speech, and Language Processing 2012

D. Dimitriadis, P. Fousek, IBM Cloud Speech to text (STT) "Developing on-line speaker diarization system" Interspeech 2017

Yanick Lukic, Carlo Vogt, Oliver Dürr, Thilo Stadelmann, "Speaker identification and clustering using convolutional neural networks" IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2016

Papers on LSTMs and d-vectors:

Quan Wang ; Carlton Downey ; Li Wan ; Philip Andrew Mansfield ; Ignacio Lopz Moreno, "Speaker Diarization with LSTM" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018

Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, Chong Wang, "Fully Supervised Speaker Diarization" ICASSP 2019

Speaker Diarization toolkit:

http://www-lium.univ-lemans.fr/diarization/doku.php/welcome and https://cmusphinx.github.io/wiki/speakerdiarization/

2.5 LANGUAGE IDENTIFICATION (LID)

Normally when using ASR, the user has to choose the language before using the ASR system, thereby allowing the system to choose a model trained on the chosen language. Good ASR systems can detect the language spoken in very short time and can automatically switch models. Some End to End systems (as shown at the end of this chapter, can be trained on multiple languages or dialects, making language identification obsolete and even more, allowing the recognition of multi-language speech.

For the best language identification decision, following information sources should/can be considered, while it can be sufficient to use only one or two.
Acoustic Phonetics. Phonetic inventories differ from language to language. Even when languages have identical phones, the frequencies of occurrence of phones differ across languages.
Prosodics. Languages vary in terms of the duration of phones, speech rate and the intonation (pitch contour). Mandarin and Vietnamese have very different intonation characteristics than stress languages such as English.
Phonotactics. Phonotactics refer to combination rules of the different phones in a language. Phonotactic rules vary across languages.
Vocabulary. Finally the most important difference between languages are the different vocabularies.

Conventional language identification models are also based on factor analysis and the corresponding i-vectors. Quickly after the manifestation of i-vectors in the field of speaker recognition (see chapter 2.4) around 2011, researchers adapted the methodology to language recognition. The models are trained using many utterances in different languages.

While classic models used i-vectors in combination with acoustic features (e.g. MFCCs), newer models using i-vectors with so called bottleneck features have managed to outperform the classic combination. Bottleneck features are extracted straight out of hidden bottleneck layer of the DNN (see Figure 6 and Figure 7).

Bing Jiang et al. introduced a bottleneck LID system in 2014 making Deep Bottleneck Features (DBF) state of the art: "By fusing the output of phonotactic and acoustic approaches, we achieve an EER of 1.08%, 1.89% and 7.01% for 30 s, 10 s and 3 s test utterances respectively." (Jiang *et al.*, 2014)

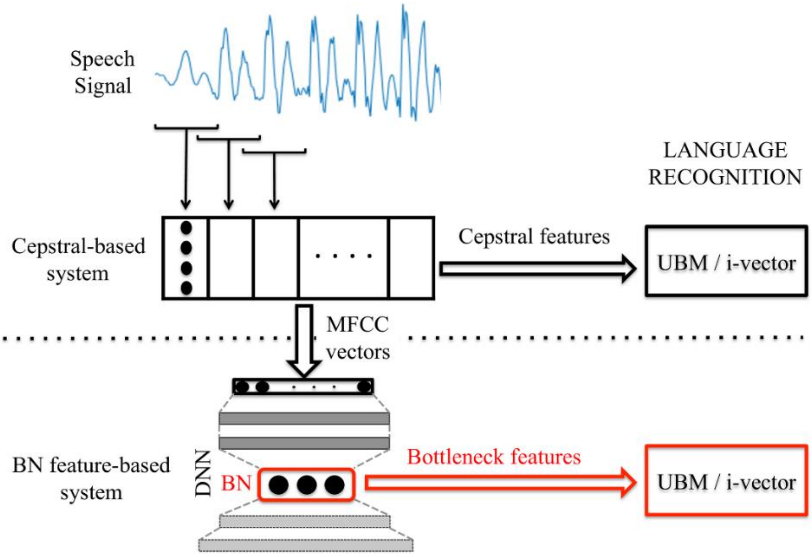


Figure 6, Representation of language recognition system structure. (Lozano-Diez *et al.*, 2017)

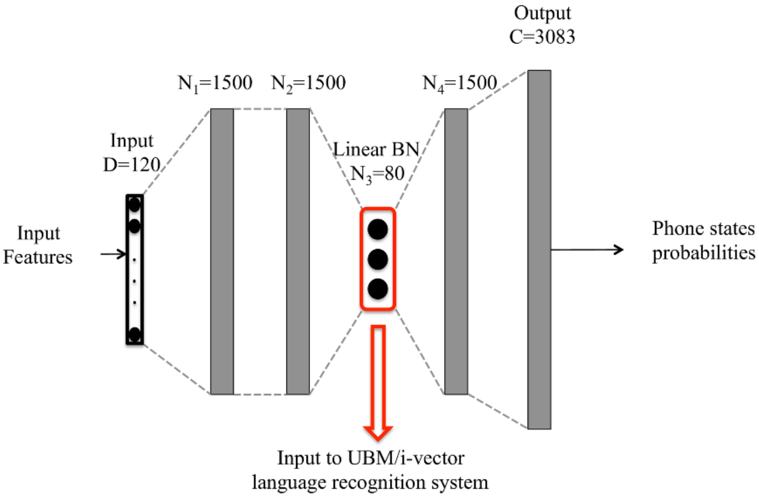


Figure 7, Bottleneck feature extraction (Lozano-Diez *et al.*, 2017)

Alternative approaches made use of LSTMs and managed to outperform conventional networks, particularly with very short utterances (~3s).

November 2016 Ignacio Lopez-Moreno et al. produced the comparison of methodologies seen in Figure 8. They compared a conventional i-vector model with a model using solely bottleneck feature vectors and a with a 'fusion' model that uses both bottleneck features and i-vector space. While the 'fusion' model is basically the state of the art bottleneck system mentioned above.

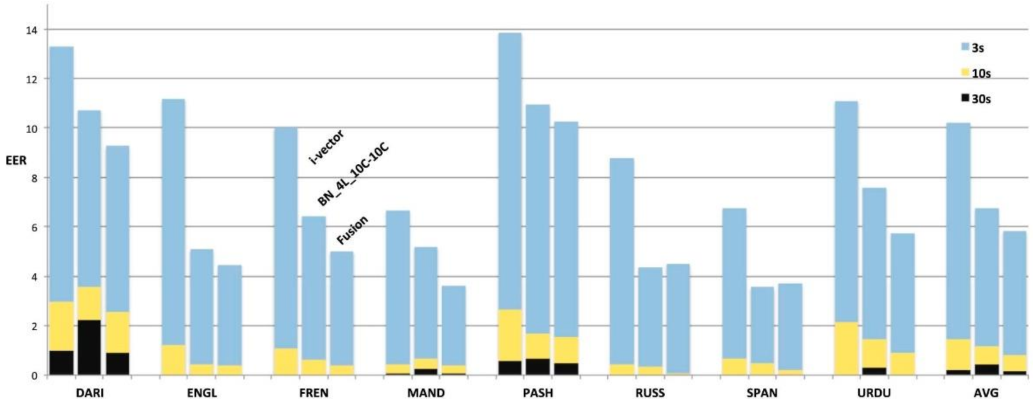


Figure 8, i-vector, state of the art bottleneck system (called fusion here) and bottleneck without i-vector system comparison (average EER) per language (Lopez-Moreno *et al.*, 2016)

Even though the i-vector approach has proven to be successful in several scenarios, there are two major disadvantages additional to the points mentioned in chapter 2.4. First, the point estimation has a larger variance when the amount of data used to compute it decreases (i.e. with shorter utterances), quickly degrading its robustness. Second, the i-vector is a compact representation of a whole utterance, so most of the computation is performed after completion of the utterance, introducing a significant latency.

New promising (alternative) approaches also address a further challenge involving mixed language in ASR (MLASR). The use of mixed language in day to day spoken speech is becoming common and is being accepted as being syntactically correct. This is a challenge to a ASR system, since in order for the system to correctly recognise, either the language of each word has to be detected individually (not practical, due to very short utterances) or the system shouldn't differentiate between the languages in the first place (see next section).

Referring to an LAS End-to-End ASR Architecture they introduced in 2017, Google sees potential for language recognition as well as applications including multi-lingual/MLASR recognition.
"Another exciting potential application for this research is multi-dialect and multi-lingual systems, where the simplicity of optimizing a single neural network makes such a model very attractive. Here data for all dialects/languages can be combined to train one network, without the need for a separate AM, PM and LM for each dialect/language. We find that these models work well on 7 english dialects [...] and 9 Indian languages [...], while outperforming a model trained separately on each individual language/dialect." (Sainath and Wu, 2017)

Papers on language identification:

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, David Martinez, Oldřich Plchot, Joaquin Gonzalez-Rodriguez, Pedro J. Moreno, "On the use of deep feedforward neural networks for automatic language identification" 2016

Bing Jiang, Yan Song , Si Wei, Jun-Hua Liu, Ian Vince McLoughlin, Li-Rong Dai, "Deep Bottleneck Features for Spoken Language Identification" 2014

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez, Pedro J. Moreno, "Automatic Language Identification Using Long Short-Term Memory Recurrent Neural Networks" Interspeech 2014

Ruben Zazo , Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, Joaquin Gonzalez-Rodriguez, "Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks" 2016

Alicia Lozano-Diez, Ruben Zazo, Doroteo T. Toledano, Joaquin Gonzalez-Rodriguez, "An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition" 2017

Alicia Lozano-Diez ; Oldřich Plchot ; Pavel Matejka ; Joaquin Gonzalez-Rodriguez, "DNN Based Embeddings for Language Recognition" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018

S. Toshniwal, T.N. Sainath, R.J. Weiss, B. Li, P. Moreno, E. Weinstein and K. Rao, "End-to-End Multilingual Speech Recognition using Encoder-Decoder Models" ICASSP 2018

B. Li, T.N. Sainath, K. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu and K. Rao, "Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model" ICASSP 2018.

2.6 VOCABULARY ENHANCEMENT WITH DOMAIN SPECIFIC NAMED ENTITIES

Defining domain specific named entities helps to ensure the correct treatment of those throughout the text.
For example, in “Activia is yogurt, but not just any yogurt” Activia is seen as a proper noun and may be recognized differently in various occasions (e.g. Activja, Activiea etc.)

To prevent misinterpretation and misclassification, the user can be permitted to define domain specific entities. In some cases the system can automatically recognize and define nouns as a domain specific entity, by querying lexica’s or such. (Brun and Hagege, 2011) propagate a system that uses WordNet to find named entity types of similar nouns. In this example with Activia, the system would assign the unambiguous type “FOOD” to Activia, since yogurt also has the type “FOOD”.

Modern End-to-End models present a challenge, since they are more successful without a language model. Language models can be used for specializing the speech recognition model according to context (user, geography, application etc.), as well as for enhancing the vocabulary with domain specific words. Baidu researchers tackle this challenge with a method called Cold Fusion introduced in August 2017 (Sriram *et al.*, 2017).

Papers on vocabulary enhancement:

T.N. Sainath, R. Prabhavalkar, S. Kumar, S. Lee, A. Kannan, D. Rybach, V. Schogol, P. Nguyen, B. Li, Y. Wu, Z. Chen and C.C. Chiu, “No Need for a Lexicon? Evaluating the Value of the Pronunciation Lexica in End-to-End Models,” ICASSP 2018.

2.7 CONVERSATIONAL ANALYSIS

In cases where ‘richer’ transcriptions are needed, Conversational Analysis (CA) is the way to go. CA is used for interactions, as they occur in doctors' offices, courts, law enforcement, helplines or educational settings. CA extracts more information from the audio signal, with the aim to describe a conversation or dialog as close to reality as possible.

To make CA possible, the National Institute of Standards and Technology (NIST) proposes the extraction of additional metadata, when performing ASR, namely edit disfluencies (revisions, repetitions, restarts), pause fillers and discourse markers.

3 Conclusion

This survey gives a solid introduction into transcription post-processing methods, while recognizing, that in some cases, as with End-to-End systems, these processes have been merged into models (e.g. language model), reducing complexity and increasing performance (providing enough training data is available).

After explaining the build of a traditional ASR system, necessary post-processing steps are explained, including recent developments. The analysis shows, that for restoring punctuation the old grammar rules have been replaced by LSTM neural networks. The network is now trained with labeled punctuation symbols and ideally with support of part of speech tagging for better accuracy, providing a level of certainty for predictions. In the field of speaker diarization, since 2011 low dimensional spaces are used to model speaker variability. The corresponding i-vector based audio embedding technique was state of the art for some years, but now have been surpassed by neural network (LSTM)-based d-vectors.
In language recognition, researchers adapted the i-vector methodology in combination with MFCCs. The models are then trained using many utterances in different languages, and with varying utterance length. However meanwhile, the state-of-the-art method, is to use i-vectors in combination with feature vectors straitly extracted out of a deep neural net bottleneck layer. Also, new LAS End-to-End systems are showing promising results, not only for language recognition, but also for mixed language in ASR.
In vocabulary enhancement with domain specific entities, the language model is enhanced with words, that are special for the domain at hand, and may not be stored in the language model. The modern End-to-End models face a challenge in this domain, since they don't have language models and require alternative methods. Baidu has therefore recently introduced a 'cold fusion' method as their solution for End-to-End networks, showing promising results.

The rapid development in speech recognition and machine learning in general, is constantly leading to new methods and opportunities and is likely going to further improve the methods mentioned above in the next few years.

4 Appendix

4.1 GERMAN IN ASR

Recognizing German language has some additional challenges, which should be considered when building a ASR model. E.g. Compounds, 7% of the tokens and 47% of the types in a 28-million token German newswire corpus are compounds. ((Marco Baroni, Johannes Matiasek, 2002) cited in (Sugisaki and Tuggener, 2018))

4.2 BIG PLAYERS IN ASR RESEARCH

- Google
- Baidu
- IBM
- Stanford University (Andrew Ng)
- Baidu's Beijing campus (Andrew Ng)
- Conell University (although, it isn't clear for me if all papers where produced in cooperation with the University)

4.3 FURTHER TOPICS IN ASR

Speech to text research topics aren't the focus of this survey. Nevertheless, an overview of current research topics is provided, based on accepted papers of the Interspeech conference.

4.3.1 Key research topics:

- Speaker Recognition
- Speaker Adaption
- Speaker Identification *(is the process of identifying the speaker from a given utterance by comparing the voice biometrics of the utterance with those utterance models stored beforehand.)*
- Speaker Verification
- Speaker Characterization
- Emotion Recognition
- Identifying Acoustic Scenes and Rare Events
- Low Resource Speech Recognition
- Dereverberation (filter noise)
- ASR for difficult languages: Indian, Mandarin, Japanese
- Recognizing Overlapped Speech in Meetings
- Application of ASR in Medical Practice (forecast and diagnose sicknesses)
- Voice activity detection
- Language Identification
- Dialectal Variation
- MLASR Mixed language in automatic speech recognition
- Secure and Private Speech Processing

While researchers are focussing on the topics above, it is worth mentioning, that the privacy aspect in handling speech data has never been this present and controversial.

4.3.2 Language modelling toolkits

- CSLM – Free toolkit for feedforward neural language models
- DALM – Fast, Free software for language model queries
- IRSTLM – Free software for language modeling
- Kylm (Kyoto Language Modeling Toolkit) – Free language modeling toolkit in Java
- KenLM – Fast, Free software for language modeling
- LMSharp – Free language model toolkit for Kneser–Ney-smoothed *n*-gram models and recurrent neural network models
- MITLM – MIT Language Modeling toolkit. Free software
- NPLM – Free toolkit for feedforward neural language models
- OpenGrm NGram library – Free software for language modeling. Built on OpenFst.
- OxLM – Free toolkit for feedforward neural language models
- Positional Language Model
- RandLM – Free software for randomised language modeling
- RNNLM – Free recurrent neural network language model toolkit
- SRILM – Proprietary software for language modeling
- VariKN – Free software for creating, growing and pruning Kneser-Ney smoothed *n*-gram models.
- Attila - IBM's Attila speech recognition toolkit

Source: https://en.wikipedia.org/wiki/Language_model#cite_ref-bengio_6-1

4.4 CONFERENCES

There are many conferences that have content relevant to the field of ASR. Below a short overview of the most important is given. Next to the Converence name, the ERA conference ranking is provided, the rankings range from A (=best) to C (=worst).

Interspeech (A-Level)

“Interspeech is the world’s largest and most comprehensive conference on the science and technology of spoken language processing”(Interspeech, 2018)

“Interspeech conferences emphasize interdisciplinary approaches addressing all aspects of speech science and technology: fundamental theories, advanced applications including computational modelling and technology development inspired by recent advances in artificial intelligence (AI) and machine learning (ML).“(Interspeech, 2018)

EMNLP (A-Level)

Conference on Empirical Methods in Natural Language Processing.

ACL (A-Level)

Annual Meeting of the Association for Computational Linguistics

ACL is the premier conference of the field of computational linguistics, covering a broad spectrum of diverse research areas that are concerned with computational approaches to natural language.

CoNLL (A-Level)

The SIGNLL Conference (ACL's Special Interest Group on Natural Language Learning) on Computational Natural Language Learning. This year, CoNLL will be colocated with EMNLP 2019. Offers insight in Attention based models.

EACL (A-Level)

European Chapter of the Association for Computational Linguistics (see ACL conference).

COLING (A-Level)

The International Committee on Computational Linguistics (ICCL) organises the International Conference on Computational Linguistics (COLING).

ICML (A-Level)

International Conference on Machine Learning

ICASSP (A-Level)

IEEE International Conference on Acoustics, Speech and Signal Processing

LREC (A-Level)

Language Resources and Evaluation

NAACL (A-Level)

North American Chapter of the Association for Computational Linguistics

INLG (B-Level)

International Conference on Natural Language Generation.

SpeechTek (No Rating found)

“Speech and conversational technologies—from voice-only IVR to visual IVR, multimodal to omnichanel, from simple commands to advanced natural language, and from smart apps to digital assistants (both voice and text)—help businesses connect with customers. These technologies help customers search, query, interact, and perform transactions easily and effectively.”(SpeechTek, 2018)

NABshow (No Rating found)

“Big-picture insights and critical details on the latest disruptive tech and trends.”(NABSHOW, 2018)

“NAB Show is where ground-breaking technology is unveiled, innovative solutions are displayed and game-changing trends are exposed. Prepare to explore aisle after aisle of awesome tech, cool gear, smart software, capable cloud solutions and limitless ideas and inspiration. Only here can you roll-up your sleeves and be hands-on with the products, services and people driving the future of content.”(NABSHOW, 2018)

5 Sources

Alibaba Cloud iDST (2016) 'Man VS Machine: The Secrets Behind Alibaba Cloud's Speech Recognition Technology', *Alibaba cloud forum*. Available at: https://www.alibabacloud.com/forum/read-183.

Amodei, D. *et al.* (2015) 'Deep Speech 2: End-to-End Speech Recognition in English and Mandarin'. Available at: http://arxiv.org/abs/1512.02595 (Accessed: 4 January 2019).

Beaufays, F. (2015) *The neural networks behind Google Voice transcription*, *Google AI Blog*. Available at: https://ai.googleblog.com/2015/08/the-neural-networks-behind-google-voice.html (Accessed: 3 January 2019).

Brun, C. and Hagege, C. (2011) 'Semantic compatibility checking for automatic correction and discovery of named entities'. Available at: https://patents.google.com/patent/US8000956B2/en.

Che, X. *et al.* (2016) 'Sentence Boundary Detection Based on Parallel Lexical and Acoustic Models', in *INTERSPEECH 2016*.

Eugenio, C. (2018) 'The fall of RNN / LSTM', *Towards Data Science*. Available at: https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0 (Accessed: 2 January 2019).

Interspeech (2018) 'Interspeech converence'. Available at: http://interspeech2018.org/.

Jiang, B. *et al.* (2014) 'Deep Bottleneck Features for Spoken Language Identification', *PLoS ONE*. Edited by D. A. Robin. Public Library of Science, 9(7), p. e100795. doi: 10.1371/journal.pone.0100795.

Juin, C. C. *et al.* (2017) 'Punctuation prediction using a bidirectional recurrent neural network with part-of-speech tagging', in *TENCON 2017 - 2017 IEEE Region 10 Conference*. IEEE, pp. 1806–1811. doi: 10.1109/TENCON.2017.8228151.

'Language model' (no date) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Language_model#cite_ref-bengio_6-1.

Levy, T., Silber-Varod, V. and Moyal, A. (2012) 'The effect of pitch, intensity and pause duration in punctuation detection', in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, pp. 1–4. doi: 10.1109/EEEI.2012.6376934.

Linagora (2018) 'Remedial course on speech recognition', *Medium*. Available at: https://medium.com/linagora-engineering/openpaas-newsletter-april-2018-888482f0e08e.

Lopez-Moreno, I. *et al.* (2016) 'On the use of deep feedforward neural networks for automatic language identification', *Computer Speech & Language*, 40, pp. 46–59. doi: 10.1016/j.csl.2016.03.001.

Lozano-Diez, A. *et al.* (2017) 'An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition', *PLOS ONE*. Edited by J. Tu. Public Library of Science, 12(8), p. e0182580. doi: 10.1371/journal.pone.0182580.

Lukic, Y. *et al.* (2016) 'Speaker identification and clustering using convolutional neural networks', in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, pp. 1–6. doi: 10.1109/MLSP.2016.7738816.

Marco Baroni, Johannes Matiasek, and H. T. (2002) 'Predicting the components of German nominal compounds.', in *15th European Conference on Artificial Intelligence (ECAI'02)*.

NABSHOW (2018) 'NABSHOW'. Available at: https://www.nabshow.com/about-nab-show/show-overview.

Pundak, G. and Sainath, T. N. (2016) 'Lower Frame Rate Neural Network Acoustic Models', in *Interspeech, 2016*. Available at: http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45555.pdf.

Sainath, T. N. and Wu, Y. (2017) 'Improving End-to-End Models For Speech Recognition', *Google AI Blog*. Available at: https://ai.googleblog.com/2017/12/improving-end-to-end-models-for-speech.html (Accessed: 4 January 2019).

SpeechTek (2018) 'Focus Areas of SpeechTek'. Washington DC. Available at: http://www.speechtek.com/2018/default.aspx.

Sriram, A. *et al.* (2017) 'Cold Fusion: Training Seq2Seq Models Together with Language Models'. Available at: http://arxiv.org/abs/1708.06426 (Accessed: 4 January 2019).

Su, H. (2018) *Combining Speech and Speaker Recognition-A Joint Modeling Approach*. Berkeley. Available at: https://escholarship.org/uc/item/9r32d8c9 (Accessed: 2 January 2019).

Sugisaki, K. and Tuggener, D. (2018) 'German Compound Splitting Using the Compound Productivity of Morphemes', in *The Conference on Natural Language Processing*.

Tilk, O. and Alumäe, T. (2016) 'Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration', in, pp. 3047–3051. doi: 10.21437/Interspeech.2016-1517.

Wang, C. (2018) 'Accurate Online Speaker Diarization with Supervised Learning', *Google AI Blog*. Available at: https://ai.googleblog.com/2018/11/accurate-online-speaker-diarization.html (Accessed: 10 January 2019).

Wang, F. *et al.* (2018) 'Self-Attention Based Network for Punctuation Restoration', in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 2803–2808. doi: 10.1109/ICPR.2018.8545470.

Wang, Q. *et al.* (2018) 'Speaker Diarization with LSTM', in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5239–5243. doi: 10.1109/ICASSP.2018.8462628.

6 List of figures