Discourse on ASR Measurement: Introducing the ARPOCA Assessment Tool

Megan Merz

Rose-Hulman Institute of Technology

Introduction

Driving Question: How can automatic speech recognition models be improved?

Key Terms:

- ASR: Automated Speech Recognition, e.g. Siri or Alexa
- WER: Word Error Rate, which measures insertions, deletions, and substitutions made by an ASR model

$$\frac{\text{minimum number of errors (i. e., number of substitutions + deletions + insertions)}}{\text{actual number of words in the reference transcript}}$$

- Corpus: A collection of sound recordings used to train an ASR model

Current Problems with ASR:

- Models with a large corpus typically score better
- The larger the corpus, the more expensive
- WER gives a score, but no insight on problems
- Models work better with some accents than others

Methodology

ARPOCA stands for Assessment of ASR Using Phonemes, Originality, Cost, and Accent Performance

- APROCA is not an ASR model itself
- Phoneme recognition pinpoints errors in recognition
- Phoneme recognition also assesses performance with different accents
- Lower cost gets better score to encourage optimization
- Originality rewards new ideas and implementations
- ARPOCA is a tool to give feedback upon ASR models
- Can be used in conferences and reviews

Originality Analysis

Models can be one of the following:

- Addition or modification of an existing model
- Original implementation using existing methods
- First implementation of an entirely new method

ARPOCA is designed to encourage originality, so entirely new methods will score highest, while additions/modifications will score lowest.

Identification of Vowels

phoneme	duration	F1	F2	F3
k	0.02977946	912.296354	2264.74853	2878.57805
e	0.1546442	745.407795	2372.94302	2876.27715
n	0.05217272	425.170747	941.633882	2261.12141
e	0.07258274	572.287818	2438.19948	2960.63973
s	0.07127291	470.293242	2207.44326	3659.47263
e	0.06572049	484.467033	2085.54805	2963.72639
s	0.0654392	445.249308	2201.00192	2756.09294
i	0.07079071	489.142181	2575.044	3439.25338
t	0.10398982	334.052411	1999.8775	2742.24203
o	0.09315965	603.574467	1273.99339	2862.64349

- Formants are a measure of frequency modeled after human perception of sound
- F1, F2, and F3 are the first, second, and third formants
- Each vowel has a range for each formant that it generally falls within

Identification of Consonants

Identifying consonants can be more challenging than identifying vowels due to the following:

- Formants cannot be used for voiceless consontants (consonants created without the use of vocal cords)
- The greater number of consonants compared to vowels in many languages
- The varying means of producing consonants, such as plosives, fricatives, and nasals

Three measurements can aid in identifying consonants, especially voiceless consonants:

- Dispersion/Standard Deviation: how widely spread the frequency of the sound is
- Center of Gravity: where the frequency of a sound is concentrated (high frequency vs low frequency)
- Intensity: the amplitude of sounds waves, used to differentiate between methods of sound production

Table 3 The Spanish consonant phonemes

			Labial	Dental	Alveolar	Palatal	Velar
Obstruents	Voiceless	Non-fricatives	p	t		ʧ	k
		Fricatives	f	θ	s		x
	Voiced		b	d		ʝ	g
Sonorants	Liquids	Laterals			l	ʎ	
		Vibrants: Tap			ɾ		
		Vibrants: Trill			r		
	Nasals		m		n	ɲ	

Phoneme Identifier Results

start time	1.93495	1.987473	2.059252909	2.129283	2.183555488	2.274594	2.330617695	2.440914
phoneme identifier	vowel	e	voiceless fricative	e	voiceless fricative	e	voiceless non-fricative	o
speech recognizer	n	e	s	e	s	i	t	o

- General identification of consonant type due to consonant identification difficulty
- Vowels are identified, but more accuracy is needed
- Comparison with speech recognizer can highlight speech recognizer problems

Cost Analysis

Estimated compute costs from the paper Energy and Policy Considerations for Deep Learning in NLP.

Model	Hardware	Power (W)	Hours	kWh-PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

- Higher cost is correlated with higher environmental impact, due to the higher compute time and energy usage
- According to NIH, average grant size in is currently about $550,000, so $500,000 may be a good baseline for entire project cost
- Compute cost and obtaining/constructing a suitable corpus can be prohibitively expensive

Conclusion

Results

- The basic framework for ARPOCA has been established and shows promise
- The phoneme recognition system is much more difficult than originally anticipated

Future Work

- More testing to determine efficacy
- Collecting more data for all areas of ARPOCA
- Further automating the phoneme recognition system
- Promotion of ARPOCA, especially for use in reviews and conferences