INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

On the Use/Misuse of the Term 'Phoneme'

Roger K. Moore, Lucy Skidmore

Speech and Hearing Research Group, Dept. Computer Science, University of Sheffield, UK
{r.k.moore,lskidmore1}@sheffield.ac.uk

Abstract

The term 'phoneme' lies at the heart of speech science and technology, and yet it is not clear that the research community fully appreciates its meaning and implications. In particular, it is suspected that many researchers use the term in a casual sense to refer to the sounds of speech, rather than as a well defined abstract concept. If true, this means that some sections of the community may be missing an opportunity to understand and exploit the implications of this important psychological phenomenon. Here we review the correct meaning of the term 'phoneme' and report the results of an investigation into its use/misuse in the accepted papers at INTERSPEECH-2018. It is confirmed that a significant proportion of the community (i) may not be aware of the critical difference between 'phonetic' and 'phonemic' levels of description, (ii) may not fully understand the significance of 'phonemic contrast', and as a consequence, (iii) consistently misuse the term 'phoneme'. These findings are discussed, and recommendations are made as to how this situation might be mitigated.

Index Terms: phoneme, phonetics, phonology, speech units, phonemic contrast, science vs. technology

1. Introduction

The idea that speech is organised around a finite set of 'fundamental' sound units is an ancient one, and many languages have exploited this phenomenon in the development of their writing systems [1, 2]. Of course the study of such sound structures is the primary remit of the speech sciences, specifically the fields of 'phonetics' (concerned with the articulatory and acoustic correlates of speech sounds [3, 4, 5]) and 'phonology' (concerned with the organisation and usage of speech sounds in any particular language [6, 7]). Likewise, the field of 'speech technology' has eagerly embraced the idea that speech may be modelled as the composition of a small set of fundamental units, and such data structures have long been implicated in the algorithms underpinning practical large-vocabulary continuous speech recognition and text-to-speech synthesis systems [8, 9, 10, 11, 12].

Of particular interest here is the concept of a 'phoneme', commonly defined as "*the smallest unit of speech that distinguishes one word from another in a particular language*" [13, 14, 15, 16]. Such a notion is central to both phonetics and phonology, and it also plays a key role in much of speech technology[1]. However, there is a suspicion that the term 'phoneme' is often used in a casual non-scientific sense, based on a mistaken belief that it is a concrete acoustic building block rather than an abstract psychological phenomenon. In particular, it is hypothesised that many speech researchers (i) may not be aware of the critical difference between 'phonetic' and 'phonemic' levels of description, (ii) do not fully understand the significance of 'phonemic contrast', and as a consequence, (iii) consistently misuse the term 'phoneme'. If this is true, then contemporary speech science and technology may be failing to exploit a crucial aspect of spoken language behaviour, and thus be unnecessarily limiting the explanatory power of our models and the capabilities of our technological solutions.

This paper reports the results of an investigation into the use/misuse of the term 'phoneme', and offers suggestions as to how the situation might be mitigated. Section 2 reviews the definition of the term, Section 3 describes the investigation, and Section 4 discusses the implications and makes three key recommendations. Finally, Section 5 concludes with summary of the main findings.

2. The 'Phoneme'

2.1. Background

According to the pioneering phonetician Daniel Jones, the idea of the phoneme was recognised from the 1870s, but the term itself was not in general use until the beginning of the 20th century [6]. The need for such a term arose because early phoneticians had realised that acoustically distinct speech sounds were *only* perceived as different (by native listeners) *if* they signalled the difference between one word and another (in that language). Crucially, acoustically distinct speech sounds (in a language) were perceived as the *same* if they did *not* signal the difference between one word and another (in that language). In other words, the sounds listeners perceive - the 'phonemes' - are conditioned on the *meaning* of an utterance, not on a fixed set of acoustic properties. Daniel Jones thus defined a phoneme as "*a family of uttered sounds*[2] *(segmental elements of speech) in a particular language*[3] *which count for practical purposes as if they were one and the same*" [6, pp. 22].

This new understanding of the dual physical and psychological nature of speech led to the realisation that any given utterance may be transcribed using *two* levels of description: the *language-dependent* 'phonemic' level (originally referred to as 'psychophonic') and the *language-independent* 'phonetic' level (originally referred to as 'physiophonic'). It also led to the requirement for agreed phonemic and phonetic transcription conventions, the founding of the International Phonetic Association (IPA) in the late 19th century, and the establishment of the international phonetic alphabet [17, 18].

As is now well established, the convention is that a *phonemic* transcription consists of IPA symbols between forward slashes, and a *phonetic* transcription consists of IPA symbols between square brackets[4]. For example, the English phrase

[1] Of course the emergence of 'end-to-end' systems raises interesting issues about the role and value of explicit intermediate levels of representation, and these are discussed in Section 4.1.

[2] Subsequently termed 'allophones'.

[3] Jones clarified that in referring to 'language', he should really use the term 'idiolect', i.e. language as used by a particular individual.

[4] It should be noted that the use of the same IPA symbol set for both phonemic and phonetic transcriptions has long been a potential source of confusion between the two for naive users.

http://dx.doi.org/10.21437/Interspeech.2019-2711

"*law and order*" could be transcribed phonemically as /lɔː ænd ɔːdɜ/, but a particular utterance could be transcribed phonetically as [lɔːrənɔːdə], where various phonological processes (e.g. elisions, assimilations, epentheses and reductions) explain the relationship between the two [19].

2.2. Implications

There are many implications that arise from the phonemic structure of spoken language. There is not space here to review the entire field, but two psychological consequences are worthy of mention. First, individual speech sounds may be 'heard'[5] even though they are not present, and second, whole words or phrases may be 'heard' even though they are not present!

The first of these is known as the 'phoneme restoration effect' [20]. Richard Warren showed that if a short section of speech was cut out and replaced by another sound (such as a cough), listeners could not detect that anything was missing; the excised sound was *restored* in the mind of the listener.

The second effect is illustrated nicely by the phonetician Sara Hawkins [21]. On hearing a verbal enquiry from a family member as to the whereabouts of some mislaid object, the interlocutor might reply with any of the following utterances:

[aɪ dəʊnt nəʊ]
[aɪ dʊnəʊ]
[dʊnə]
[ə̃ə̃ə̃]

... where the last utterance is barely more than a series of nasal grunts! Which of these utterances is actually spoken would depend on the communicative context. Indeed, the example illustrates how speakers and listeners continuously balance the effectiveness of communication against the *effort* required to communicate effectively [22] - behaviour that leads to a 'contrastive'[6] (as opposed to signal-based) form of communication [23]. However, the point here is that the listener perceives /aɪ dəʊnt nəʊ/ ("*I don't know*") in each case! Likewise, since it is *top-down* context that facilitates the appropriate perception, the utterance [ə̃ə̃ə̃] might be easily perceived as a completely different sequence of words in a different scenario.

Other significant phenomena include 'coarticulation' which, contrary to what most speech technologists think, is not just a local effect, but which can span entire syllables [19]. Indeed, the spread of phonemic information over time (due to asynchronous control of the articulators) means that, almost by definition, there are *no* acoustic boundaries between phonemes (thereby rendering any 'beads-on-a-string' assumptions fundamentally flawed [24]).

Of course, all of the above should be familiar to anyone working in speech research. However ...

3. The Study

In order to gauge the usage of the term 'phoneme' in the broad speech science and technology community, it was decided to analyse the texts of all papers accepted for publication at the most recent INTERSPEECH conference - INTERSPEECH-2018 - which took place in Hyderabad, India in August 2018. 791 papers comprising a total of over 3 million words were analysed, and it was found that 34% of the papers contained at least one occurrence of the word "*phoneme*", with an average of 7.69 occurrences per paper[7].

[5] I.e. 'perceived'.
[6] I.e. 'discriminative'.
[7] One paper contained 88 occurrences of the term 'phoneme'!

To put these figures into context, Table 1 shows the statistics for various other keywords occurring in the INTERSPEECH-2018 accepted papers. As can be seen, "*phoneme*" occurred more frequently than "*speech synthesis*", but half as often as "*speech recognition*". Unsurprisingly, "*speech*" occurred in all papers, with an average of 44.46 occurrences per paper.

Table 1: *Usage of the terms 'phoneme', 'speech', 'speech recognition' and 'speech synthesis' in the 791 accepted papers at INTERSPEECH-2018.* ***Count*** *refers to the total number of occurrences,* ***#P*** *and* ***%P*** *refer to the number and percentage of papers in which the term appears respectively,* ***Av.*** *refers to the average number of occurrences in those papers, and* ***Max.*** *refers to the maximum number of occurrences in any one paper.*

	phoneme	*speech*	*speech rec.*	*speech synth.*
Count	2038	35171	3210	780
#P	265	791	536	152
%P	34%	100%	68%	19%
Av.	7.69	44.46	5.99	5.13
Max.	88	196	35	23

3.1. Historical Comparison

For a historical perspective, Table 2 shows the statistics for the same keywords occurring in papers accepted for the 5th International Conference on Spoken Language Processing - ICSLP-1998 - which took place in Sydney Australia, 20 years earlier than the Hyderabad INTERSPEECH. Comparing the data shown in Tables 1 and 2, it can be seen that the overall pattern of usage has not changed greatly over the intervening period (apart from the rather surprising observation that the word 'speech' was significantly less frequent in ICSLP-1998 - and was even missing from 2% of the papers!).

Table 2: *Usage of the terms 'phoneme', 'speech', 'speech recognition' and 'speech synthesis' in the 831 accepted papers at ICSLP-1998.*

	phoneme	*speech*	*speech rec.*	*speech synth.*
Count	2505	19147	2251	463
#P	290	816	482	140
%P	35%	98%	58%	17%
Av.	8.64	23.46	4.67	3.31
Max.	75	121	29	20

Figure 1 shows the distribution of the occurrences of the term 'phoneme' in the accepted papers for the two conferences. As can be seen, the distributions are quite similar; the term did not appear at all in 66% of the INTERSPEECH-18 papers and 65% of the ICSLP-98 papers.

Both data sets were also analysed to determine the most frequent words to occur immediately before or immediately after the word "*phoneme*". The results are shown in Table 3. As can be seen, there is a reasonable amount of agreement, with similarly high occurrences of "*phoneme based*", "*phoneme recognition*" and "*phoneme sequence*" in both conferences. However, what is particularly noticeable is that most of these frequently occurring bigram phrases are more related to speech technology than to speech science.

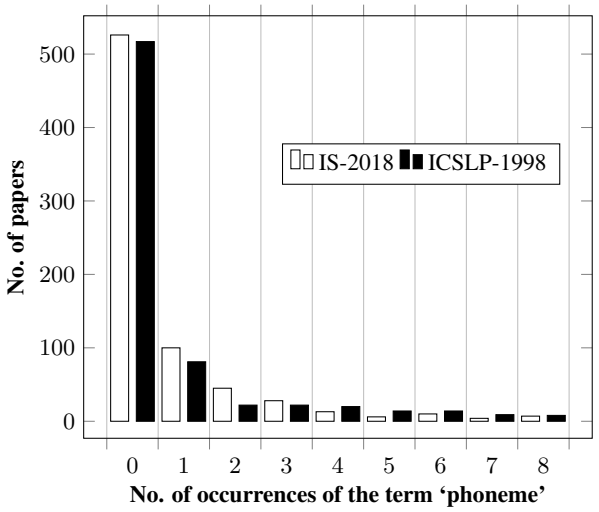

Figure 1: *Distribution of the occurrences of the term 'phoneme' in the INTERSPEECH-2018 and ICSLP-1998 accepted papers.*

Table 3: *List of the most frequent words to occur adjacent to the word "phoneme" in the accepted papers at the INTERSPEECH-2018 and ICSLP-1998 conferences.*

IS-18	ICSLP-98
based	*recognition*
sequence	*similarity*
recognition	*model*
label	*sequence*
level	*based*
conversion	*boundary*
set	*duration*
CI	*string*
classifier	*class*
classification	*dependent*

3.2. Accuracy of Definitions

One potentially clear indicator of the appropriate or inappropriate use of the term 'phoneme' is to appraise the accuracy/adequacy of any provided definitions. However, of the 265 accepted papers in INTERSPEECH-2018 that contained "*phoneme*", only six had anything approaching a definition and, of those, none were satisfactory[8]. For example, papers contained statements such as "*Speech signal consists of various basic speech sound units, which are called as phonemes.*" and "*. . . treat any sub-word acoustic unit as a phoneme.*". One pseudo-definition was simply incorrect: "*In phonetics, it is believed that when one pronounces two neighbouring phonemes, there often exists joint frames that can be a very short pause belonging to neither phoneme . . .* ". The remaining efforts superficially equated phonemes with "*sound symbols*" or with "*sub words*". Interestingly, of these six attempts at a definition, five were in papers presented in speech technology (as opposed to speech science) sessions.

However, what is perhaps most concerning is that 98% of the INTERSPEECH-2018 papers that used the term 'phoneme' did not provide any form of definition or explanation, presumably because the authors assumed that everyone knows what it means.

3.3. Use and Misuse

Further analysis of the 265 accepted papers in INTERSPEECH-2018 that contained the term 'phoneme' showed that around 40% used the term in a way that could be construed as misuse. For example, some authors seemed not to be aware of the crucial difference between phonemic and phonetic levels of description and the associated /. . . / versus [. . .] convention for transcription. Other authors clearly assumed that there is a fixed relationship between a 'phoneme' and its acoustic realisation, and that boundaries between 'phonemic segments' existed and were well defined. As a consequence, many authors referred to 'phonemic segmentation' and the derivation of 'phoneme durations'. One even referred to "*phoneme chunks*"!

Other examples of misuse include the following:

- "*We have 252 phonemes, of which there are 213 Mandarin and 39 English.*" – illustrates a fundamental misunderstanding of how phonemes are defined and enumerated.
- "*There are 144 traditional phoneme states in a mono phone system.*" – reveals a confusion between levels of representation.
- "*. . . a set of isolated phonemes extracted from CS [continuous speech] sentences.*" – shows a lack of understanding of articulatory dynamics.
- "*. . . context-dependent phonemes.*" – a lack of specificity as to what level of representation is being modelled.
- "*. . . treating filled pause as a special 'phoneme'.* " – demonstrating a cavalier application of the term to a non-linguistic event.
- "*Spectral transitions between phonemes . . .* " – false assumption that phonemes are acoustic units.
- "*There is no clear interpretation of HMM states for emotion recognition as for automatic speech recognition (sub phoneme).*" – casual interpretation of different types of representation.
- "*We propose a language-independent phoneme segmentation method.*" – shows a lack of understanding of the essential language-specific nature of phonemes.
- "*Even though they all use the same phoneme symbols, each language and accent imposes its own coloring or 'twang'.*" – overly informal description of phonetic variability.
- "*. . . modeling only the correct pronunciation of each individual phoneme.*" – gross assumption about phonetic variation.
- "*we consider the AUs [acoustic units] as phonemes* " – false assumption that phonemes are acoustic units.
- "*. . . fMLLR normalized features which are speaker independent phoneme representations.*": false assumption that phonemes are acoustic units.
- "*. . . If a phoneme lasts for more than 5ms . . .* " – false assumption that phonemes are acoustic units.
- "*. . . below the minimum duration of a phoneme (30 ms) are considered as spurious regions.*" – ignoring the realities of speech production.
- "*Diphthongs and triphtongs [sic] are split into their constituent phones to reduce the number, and enforce sharing, of phonemes.*" – failure to appreciate that diphthongs and triphthongs may be phonemes themselves.
- "*. . . universal phoneme mapping . . .* " – gross assumption about the relationship between the sounds in different languages.

[8]Note that citations to individual papers will not be provided in order not to embarrass the author(s).

- "*At a local, temporally constrained level, we observe concrete linguistic events (phonemes) . . .*" – misunderstanding about the abstract psychological nature of phonemes.
- "*. . . trained with context-independent phoneme states as targets.*" – false assumption that phonemes are acoustic units.

Of course, it should be acknowledged that around 60% of the 265 accepted papers in INTERSPEECH-2018 that contained the term 'phoneme' did *not* misuse the term in an inappropriate manner. In particular, it was noticeable that authors in the areas of L2 learning and low-resource languages were considerably more precise in their usage. However, most other mentions were single occurrences where the term 'phoneme' was used in a casual/generic sense, e.g. to refer to a category label in a classifier.

3.4. Science vs. Technology

As mentioned above, it is conceivable that the term 'phoneme' is used differently in in different parts of the research community. In order to test this, all of the accepted papers in INTERSPEECH-2018 were categorised according to whether they fell into the speech science or speech technology areas. This resulted in 185 papers tagged as 'science' and 606 papers tagged as 'technology'. Figure 2 shows the distribution of the occurrences of the term 'phoneme' based on this categorisation. As might be expected, there is evidence that the term 'phoneme' occurs slightly more frequently in the speech science papers. However, it was found that the term did not appear at all in 67% of the speech science papers and 66% of the speech technology papers - remarkably similar proportions.

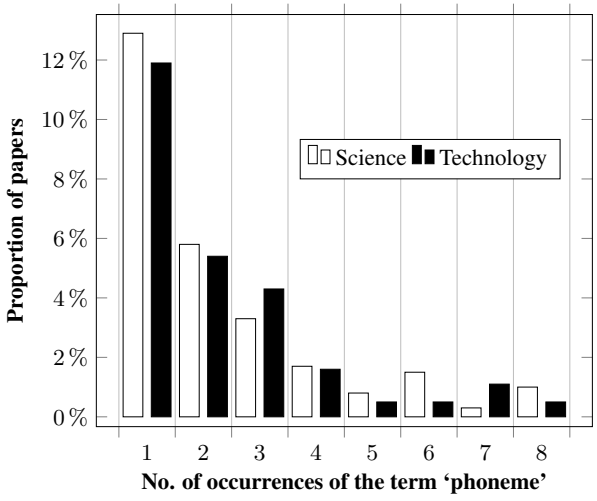

Figure 2: *Distribution of the occurrences of the term 'phoneme' in the INTERSPEECH-2018 accepted papers in the speech science and speech technology categories.*

Returning to the discussion of use/misuse in Section 3.3, it turned out that, of the papers mentioning the term 'phoneme' in potentially inappropriate ways, 25% were categorised as 'science' and 75% as 'technology'. However, since there were approximately three times as many speech technology papers than speech science papers, it seems that the accuracy associated with using the term is more or less the same in the two sections of the community; a somewhat surprising result.

4. Discussion and Recommendations

The results of this investigation clearly demonstrate that, although the term 'phoneme' is used quite frequently by the speech science and technology community, it is often deployed in a casual informal manner without considering its deeper formal implications. This means that, in many cases, the term 'phoneme' could have been substituted by 'phone' with no loss of meaning. Of course, part of the reason for this situation is that many language resources are supplied with predefined so-called 'phonemic' labels, often using their own non-IPA symbol sets. Clearly, this encourages users of such resources not to question the nature and significance of such labels. In particular, using forced-alignment with such labels reinforces the superficial 'beads-on-a-string' view of speech (c.f. Section 2.2).

4.1. Why it Matters

A reader might be forgiven for asking whether the issues raised in this paper have any great significance for future research in speech science and technology. Indeed, it is acknowledged that there is discomfort within some sections of the community about the validity and usefulness of the 'phoneme' as a theoretical construct [25, 26], and some proponents of so-called 'end-to-end' systems actively reject the concept [27]. Nevertheless, the opinion of the authors is that there is a small but not insignificant risk that future research may fail to exploit an important aspect of spoken language behaviour, and thus unnecessarily limit the explanatory power of the derived models and the capabilities of technological solutions.

Using the term 'phoneme' correctly is certainly a small price to pay to avoid such outcomes, and may even lead to deeper and more valuable insights into the structure and behaviour of spoken language - especially when coupled with contemporary ideas in deep learning [28], such as 'generative adversarial networks' [29] and 'attention models' [30].

4.2. Recommendations

Based on the observations reported in this paper, it is possible to make three key recommendations . . .

1. **Researchers** should avoid the term 'phoneme' unless they are certain of its meaning. In particular, the term 'phone' should be used to describe a generic speech sound, and the term 'phoneme' should be reserved to refer to the abstract family of sounds that serve to distinguish one word from another in a particular language.
2. **Teachers/supervisors** should ensure that newcomers to the field of speech science/technology are fully briefed on the critical difference between 'phonetic' and 'phonemic' levels of description, the significance of 'phonemic contrast', and the correct usage of the term 'phoneme' [31, pp. 206].
3. **Community associations** (such as ISCA and IEEE) should take steps to ensure that their members are aware of the importance of using the term 'phoneme' correctly.

5. Summary and Conclusion

The investigation reported in this paper has confirmed the hypothesis that a significant proportion of the community (i) may not be aware of the critical difference between 'phonetic' and 'phonemic' levels of description, (ii) may not fully understand the significance of 'phonemic contrast', and as a consequence, (iii) consistently misuse the term 'phoneme'. Three key recommendations are made that aim to mitigate the situation.

6. References

[1] P. T. Daniels and W. Bright, Eds., *The World's Writing Systems*. Oxford: Oxford University Press, 1996.

[2] G. Sampson, "Writing Systems," in *The Routledge Handbook of Linguistics*, K. Allan, Ed. Abingdon: Routledge, 2016, ch. 4, pp. 47–61.

[3] P. Ladefoged, *Elements of Acoustic Phonetics*. London: University of Chicago Press, 1962.

[4] J. D. O'Connor, *Phonetics*. Harmondsworth, UK: Penguin Books, 1974.

[5] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Mass.: MIT Press, 1998.

[6] D. Jones, "The history and meaning of the term 'phoneme'," in *Phonology: Selected Readings*, E. C. Fudge, Ed. Harmondsworth, UK: Penguin Books, 1973, ch. 1, pp. 17–34.

[7] J. Bybee, *Phonology and Language Use*. Cambridge: Cambridge University Press, 2001.

[8] J. N. Holmes and W. Holmes, *Speech Synthesis and Recognition*. Taylor & Francis, 2002.

[9] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.

[10] M. Gales and S. J. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[11] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[12] R. Pieraccini, *The Voice in the Machine*. MIT Press, Cambridge, MA, 2012.

[13] "Phoneme," in *Merriam-Webster Dictionary*. [Online]. Available: https://www.merriam-webster.com/dictionary/phoneme

[14] "Phoneme," in *Collins English Dictionary, Harper Collins Publishers*. [Online]. Available: https://www.collinsdictionary.com/dictionary/english/phoneme

[15] "Phoneme," in *Encyclopedia Britannica*. [Online]. Available: https://www.britannica.com/topic/phoneme

[16] "Phoneme," in *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Phoneme

[17] "International Phonetic Association." [Online]. Available: https://www.internationalphoneticassociation.org

[18] M. K. C. MacMahon, "The International Phonetic Association: the first 100 years," *Journal of the International Phonetic Association*, vol. 16, pp. 30–38, 1986.

[19] M. Ashby and J. Maidment, *Introducing Phonetic Science*. Cambridge University Press, 2005.

[20] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, no. 3917, pp. 392–393, 1970.

[21] S. Hawkins, "Roles and representations of systematic fine phonetic detail in speech understanding," *Journal of Phonetics*, vol. 31, pp. 373–405, 2003.

[22] E. Lombard, "Le sign de l'élévation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[23] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990, pp. 403–439.

[24] M. Ostendorf, "Moving Beyond the 'Beads-On-A-String' Model of Speech," in *IEEE ASRU Workshop*. Keystone, USA: IEEE, 1999, pp. 79–84.

[25] G. Sampson, "Is there a universal phonetic alphabet?" *Language*, vol. 50, no. 2, pp. 236–259, 1974.

[26] R. F. Port and A. P. Leary, "Against formal phonology," *Language*, vol. 81, pp. 927–964, 2005.

[27] E. Shafaei-Bajestan and R. H. Baayen, "Wide learning for auditory comprehension," in *INTERSPEECH 2018*. Hyderabad, India: ISCA, 2018, pp. 966–970.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems (NIPS) Conference*, Montreal, Canada, 2014.

[30] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[31] D. Gibbon, R. K. Moore, and R. Winski, Eds., *Handbook of Standards and Resources for Spoken Language Systems*. Berlin, New York: Mouton de Gruyter, 1997.