THE THOUGHT COLLECTIVE BEHIND THIRTY YEARS OF PROGRESS IN SPEECH SYNTHESIS

Carina Lozo, Jan Luttenberger, Michael Pucher

Acoustics Research Institute, Austrian Academy of Sciences
carina.lozo@oeaw.ac.at

Abstract: Speech synthesis as a complex research area draws on several disciplines to further its overall goal, the recreation of human speech. The discourse of the field changed over time, as new ideas and possibilities were integrated into synthesis systems and models, reflecting the theoretical and technical developments and changes. In this contribution we illustrate and reflect this process by means of discourse analysis on the semantic and lexical level. With epistemic concepts adapted from Ludwik Fleck [1] we can identify knowledge as a form of social practice which produces a thought style within the *Speech Synthesis Workshop's* (SSW) thought collective. Thought styles surface on the discourse fragments produced by the thought collective manifested in their communication of thoughts. To grasp the thought style of the SSW collective we take a look at the abstracts ($N = 500$) of SSW proceedings from 1990 to 2016. Since the text type *abstract* forms a relevant part of the participation in the SSW, they have to exhibit accepted expressions of the collective's thought style in order to gain access to the communication of thoughts. Starting from key models in the field, we look at their attribution, frequency of usage and their use as categories for describing specific systems. We identify these key models as Neural Networks, Hidden Markov Models and Unit Selection speech synthesis. We expect a shift in emphasis of the key terms by the collective over time, as new models and concepts are introduced and adapted to the thought style, thus reshaping it.

Index Terms: History of speech synthesis, discourse analysis, Speech Synthesis Workshop

1 Introduction

The idea of a machine with a human-like voice sparked a whole research field in the past century, which is still characterized by a growing, versatile and interdisciplinary research community. The history of speech synthesis reflects how new findings and technological advances originating from different disciplines can cause a shift of a community's foci and desiderata.

At a time when the sound of immense data storage was still exotic, the significant upturn in computer technology bestowed new possibilities upon the research community in the mid-1980s. With growing computing capacities, now affordable to scholarly organizations, the interest in digital speech technologies outside the telecommunications industry also grew bigger. The speech community's aspiration to explore this vast potential became apparent, amongst other things, in the first *Speech Synthesis Workshop* (SSW) held in Autrans, France in 1990.

Prominent institutions were also founded during this time, such as the Centre for Speech Technology Research (CSTR) in 1984, Edinburgh; the Language Technology Institute (LTI) in 1986, Pittsburgh, or the Advanced Telecommunications Research Institute (ATR) in 1989, Kyoto.

From the start, the SSW carried out the function of connecting researchers, providing a scholarly space for new thoughts, ideas and development to be discussed and circulated within a group of not only established but also new members of the community. In order to describe this research community, we apply Fleck's concept of *thought collective*. A thought collective is constituted by a group of individuals, who are exchanging ideas or maintaining a kind of intellectual interaction with each other, which Fleck calls *communication of thoughts*. Thought collectives are not only limited to scientific collectives. To better understand the structure of a thought collective we need to introduce a significant concept beforehand. The *thought style* presents the most important characteristic of a thought collective. Fleck postulates that a thought style is directed perception with the appropriate processing of the perceived.* It also dictates common features of a collective's desiderata; facts which are indisputable to the collective or even methods which are generally applied in the field. When a thought style becomes inherent, the collective divides itself into an esoteric and exoteric circle. Whereas the esoteric circle includes subject-specific professionals, the exoteric circle includes interested laypersons [1]. In our case, people presenting at the SSW form the esoteric circle, as they are among the current experts on the speech synthesis field. These two circles are by no means isolated from each other. Since each circle exerts influence on the other, their relationship is considered dynamic. Although it is reluctant to change, a thought style is not as rigid as it might seem. There are phases when it's more open to changes. At these points, a paradigm shift can happen. As later described in detail, we can see those changes represented in the abstracts of the SSW proceedings.

Whereas the collective is not aware of their thought style, the communication of thoughts presents an essential instrument for the thought collective, since it regulates the participation of its members. Hence the active participation itself postulates an important function for the members of a collective, e.g. for the general admission to the collective or the status within the collective. It is crucial to be successful in distributing your ideas to the thought collective in order to gain recognition (via citations for example) for your work. Contributing to the communication of thoughts is an important part in claming full membership to the collective. There are many ways of rating the contributions to the communications of thoughts. Simple examples would be the higher recognition of peer-reviewed articles compared to non-peer-reviewed or *h*-index etc. They are all dimensions for the participation in the communication of thoughts.

Elements of the past, the future as well as mutualities are manifested in any kind of knowledge [1], making it thereby socially and historically conditioned. By investigating the communication of thoughts disclosed in the SSW abstracts from 1990 to 2016, we can show how these social conditions change over time and interact with the historic technical advances, hence resulting in new directions in the field.

2 Methods

We collected a text corpus of 500 abstracts submitted to the SSW from 1990 to 2016. The abstracts were analyzed with AntConc [2] to access single lexeme tokens. Index terms have been excluded since they were only introduced in SSW 6 (2007). Furthermore, information regarding number of citations, authors and authors' affiliation was collected and annotated. For the text analysis, we draw on the DIMEAN model of Spitzmüller and Warnke [3] for orientation. Starting at the intratextual level we examined frequently used single lexemes describing methods, systems and models described in the abstracts in regard to linked propositions and attributions. On the interactant level we looked at the authors and their affiliation to determine

*To illustrate this with a simple example: the internalized thought style lets a phonetician easily identify stops or fricatives in a spectrogram, whereas for laypersons a spectrogram is nothing but gray noise.

whether personal preferences can be observed and if affiliation plays a role. Finally, since our data spans over 26 years, we followed the usage of terms over time to observe the changes in thought style caused by new developments during the period.

3 Material

The writing of the examined abstracts is embedded in the conduct and practice of the SSW as a scholarly gathering and the association organizing it. We see the SSW as a focal event of the thought collective, setting a major stage for communication of thoughts and thereby allowing the members of the collective to perform various important acts of communication. Furthermore, we argue that the writing of an abstract constitutes a separate genre subservient to the whole process of conduct (of the SSW). Following Swales [4], we define genre as a class of semiotic acts serving a particular set of communicative purposes for a certain discourse community, which we equate with the thought collective of the SSW.

In our case, the abstract as part of the SSW practice negotiates who may participate in the communication of thoughts by presentation and publishing of an article. Therefore certain formal, stylistic and content-related characteristics have become mandatory features recognized by the members of the community. These features are the result of a long history of (scientific) social practice, making the evaluation of new texts a highly intertextual enterprise [5]. From an interactional viewpoint this means that some members of the thought collective take on the role of reviewer, evaluating if a submitted abstract meets the criteria of the genre. For applicants this means they have to design their abstracts accordingly (or else run the risk of being barred from participating in the communication of thoughts), thus minimizing the *intertextual gap*. At the same time, this unavoidable deviation can also be exploited as a means of distinction and, if met with sufficient acceptance and anticipation by the community, can lead to a change of practice. Since they are in fact published, it is safe to assume that the investigated abstracts have passed evaluation. Therefore, we can also assume that they are adapted to the collective's thought style to a degree that met the reviewers' expectations, allowing us to look at them as exemplary exhibitions of a particular thought style.

At this point it is also important to note that the communication of thoughts comprises many more texts published and anticipated elsewhere. Rather than being the place to negotiate ideas, practices etc. we chose the SSW abstracts as important places of demonstrating thought style before other members of the collective in an effort to claim membership.

4 Analysis

For the analysis in this paper we focus on the different concepts which were used to synthesize speech. In fact, there are many other issues which the collective was occupied with through the timespan of 1990 to 2016. Prosody related issues for example can be seen as a clear thread running through over the whole period of time.

The past three decades have seen the rise of three main concepts for synthesizing speech. By a short description, we illustrate how knowledge is historically conditioned, since features of each concept are manifested in the prospective systems. The burden of the past as Fleck states it, limits the possibilities to be explored by the collective, since new knowledge always results from existing knowledge. Thus other important concepts such as articulatory speech synthesis or formant synthesis were in fact relevant issues in the beginning of the SSW, as seen in figure 1, but do not seem to draw the SSW collective's interest in the following years.

For the purpose of better understanding, we describe the concepts *Di-/Triphone synthesis* or *Unit Selection (US), Hidden Markov Models (HMM)* and *Neural Networks (NN)* hereinafter.

In the 1970s the computing capacities allowed the concatenation of waveforms to gain more and more importance. Due to the artefacts resulting from successively concatenating whole phone units without proper computing of the transitions, diphone and triphone synthesis approaches were developed. These approaches take units (whole waveforms, stored in databanks), which range from the center of one phone to the center of the following, as a basis for the synthesis, omitting thereby the transition artefacts. However, the lack of prosody modeling results in unnatural sounding speech. When in the mid-1990s statistical approaches were able to remedy this problem, concatenative synthesis represented the state of the art until the late 1990s. It has to be noted, that over the years, although based on the same concept, the notion of "unit selection" evolved to be the dominant term and the "di-/triphone" notion was gradually omitted.

Statistical approaches entered the realm of speech synthesis early in the SSW era, most prominently in the form of HMMs. However, HMMs were first used for automatic speech recognition and later for synthesis. While applied for improving the preprocessing of speech for US at first, it was not until the early 2000s when the first HMM based synthesis systems were published. In contrast to a conventional US system, HMM systems do not need to store authentic speech signals. They are trained by a speech corpus and then they derive the statistical models for each unit and store only these models. An HMM states a probabilistic sequence model, which labels each unit of a given sequence (Markov chain). This sequence can consist of sentences, words or phones and is modelled with hidden states by the system. Since it is a probabilistic sequence model, the HMM calculates the probability distribution of the given label sequence and selects the most likely model for the speech signal.

Since 2010 (Deep) Neural Networks (DNN) slowly seem to replace HMM based speech synthesis. NNs were again first applied in automatic speech recognition and later entered the synthesis field. After training with enormous amounts of data, DNNs can recognize feature patterns of speech and by means of artificial neural nets consequently generating those patterns.

By no means these concepts can be considered as isolated methods. Combined they demonstrate how knowledge within and outside the collective has accumulated over the past century and currently peaks at the DNN approach. Figure 1 illustrates how the relevance of each concept has changed over time. While in 1990 the Di-/Triphone synthesis was popular and a desideratum for the collective, US or HMMs were only represented fractionally. After the release of the first open-source toolkits for speech synthesis in the early 2000s, the relevance of statistical approaches increased.

What may surprise is the early appearance of "neural networks". This shows how long an idea can circulate within the collective as it is getting more and more adapted to the thought style. After disappearing, NNs re-emerge suddenly in 2016. Section 4.1 describes this development in detail.

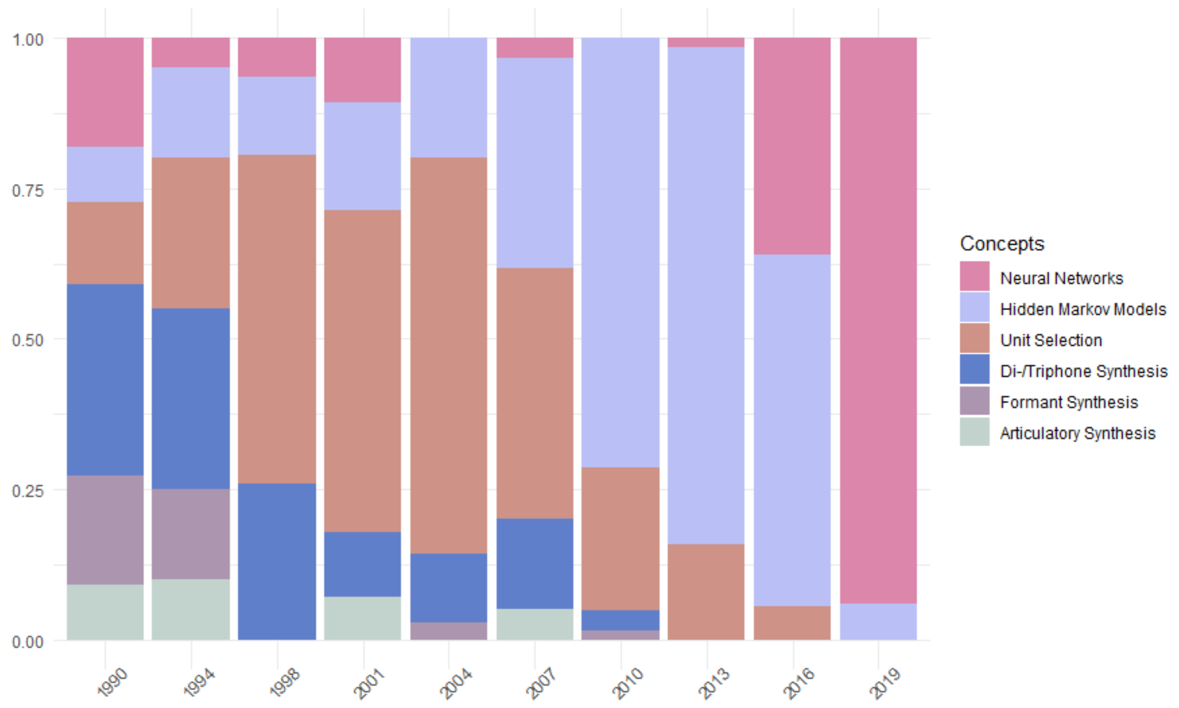

Figure 1 – Comparative usage of the synthesizing concepts over time

4.1 Neural Networks

To illustrate the ever changing focus of the thought collective, we want to take a close look at the term "neural network". While it is around since the first SSW in 1990, its meaning and importance has greatly shifted with the application of recent advancements in computational technology summarized under the name of *deep learning*. Rather than being an invention for synthesis alone, NNs have been adapted from general computer engineering for synthesis purposes, also reflecting how outside innovations change the thought collective's discourse. Across the whole corpus ($N = 500$), the term "neural network" appears in 28 abstracts (5.6%). While this number seems fairly small its distribution over time is far from even: 14 of the 28 of the submissions were submitted to SSW 9 (2016), making up almost 38% of its total contributions ($N = 37$). While papers with the term were not cited very often compared to papers with the highest absolute citation count, there are several outstanding texts. In three years papers related to NNs take the lead in citation count, namely in 1990 ([6], $N = 135$ citations), 2013 ([7], $N = 49$ citations), and 2016 ([8], $N = 98$ citations). This correlates roughly with the overall appearance of the term. Besides the mentioned three, no paper exceeded 25 citations and only one more than 15 so far. From a statistical, topical and actor-centered viewpoint, neural network's history at the SSW can be split in two parts: The early NN era from 1990 to 2007 and, after a gap in 2010, the *Deep Neural Network* era, featuring heavily in 2016. The early era has no clear center of development, except maybe for a weak tendency towards German speaking Europe, considering Traber's [6] high citation count in 1990 and contributions from Dresden University and Siemens. Also, Edinburgh University's CSTR and Southampton University take part in the early discourse, especially with CSTR being an important center in general. Topically, employing NNs for (all) parts of speech synthesis to create new systems starts in 1990 as a novelty and the attempts for new systems continue until 2007. Besides system creation, questions around pitch contours, F0 generation and boundary assignment are the main focus of the papers and a frequent field for employment of NNs. This is well in line with the general focus of the field around improving concatenative synthesis.

Things change with 2013/2016: NNs are a well-established means of computation by then. In 2016 a whole oral session with four submissions is dedicated to *Deep Learning in Speech Synthesis*, showing its acceptance and relevance. Its commonality and successful application is stated several times in 2016's abstracts. What is new is the *depth* of the networks: The adjective "deep" is always employed starting with 2013 as well as the corresponding acronym "DNN", first appearing in the contribution by Lu, Watts and King from CSTR [7]. Together with the collaborating Japanese National Institute of Informatics (NII) in Tokyo, CSTR becomes the main center in using the term. While applying DNNs to create synthesis systems is still a topic, there are several papers discussing DNN architectures and input, focusing on improving DNNs (rather than comparing them to other methods). Therefore, we can see NNs are no longer tied to individual systems, but rather have become an independent component to be examined and worked on. This is also reflected in the software *Merlin*, which is labeled a "toolkit" rather than a "system". It is also remarkable that the citation count ($N = 98$) for its presenting paper is almost ten times as high as the second most cited paper of 2016, indicating lively anticipation. One has also to consider the development of the whole field besides neural network technology alone. In 1990, statistical-parametric approaches were still in their infancy, leaving concatenative synthesis as state-of-the-art input for NNs. This changed completely in 2013 and 2016: Input as well as output for synthesis were now thought about along different lines, opening new possibilities for the application of DNNs. This also sheds an insightful light on the frequency of the term's usage: In 1990 all kinds of different approaches were brought to the fore in search for better synthesized speech. When HMM was the field of rapid advancements until it became the state-of-the-art, NN approaches were either a niche topic or gone altogether. As soon as HMM theory had reached its peak and was in need of improved computational methods, the interest in NN re-emerged suddenly from specific centers of focus. Those centers, CSTR and NII, already distinguished as leading institutions among the thought collective, were able to reintroduce the term in a way that in 2016 almost 38% of abstracts were using it, all of them now referring to "deep neural networks".

4.2 Citations as a dimension for the communication of thoughts

Since the communication of thoughts presents an essential instrument for the collective and especially for the individual members to gain recognition, a rather obvious dimension to quantify the contribution to the communication of thoughts is the number of citations. Figure 2 illustrates the information about the number of citations from `https://scholar.google.at/`, collected in early 2019. System describing papers state exceptions through out the time. As shown in figure 2, papers describing milestone systems seem to exceed other contributions in their period by far, like the US systems *German Mary TTS* [9] and *Festival* [10], the HMM toolkit *HTS 2.0* [11], as well as a paper dealing with a speech corpus for speech synthesis research [12]. Due to the open-source nature of most of these systems, they consequently gain more importance. This can also be seen in the recent SSW 9, where this tendency with the *Merlin* system [8] is also showing.

Table 1 describes the mean number of citations ($\bar{x}$), the total number of citations and total number of papers per period. To not distort the picture, we excluded outlying papers from this table ($N > 150$). Identifying and introducing relevant issues to the collective, such as US and HMMs, hence bringing them into circulation is a prerequisite before they become adapted to the thought style and presumed as indisputable at last. Based on the mean citations as Table 1 shows, we conclude that such a process took place during the period 1998 to 2004 and thus we

Table 1 – Summary of citations per period.*

Period	1990	1994	1998	2001	2004	2007	2010	2013	2016
$\bar{x}$ cites	13.4	18	23.6	24.7	29.1	11.8	7.8	9.4	3.5
total citations	927	1139	1371	1162	1312	829	475	491	130
total papers	69	63	58	47	45	70	61	52	37

*Without system describing papers.

consider this time highly formative for the collective.

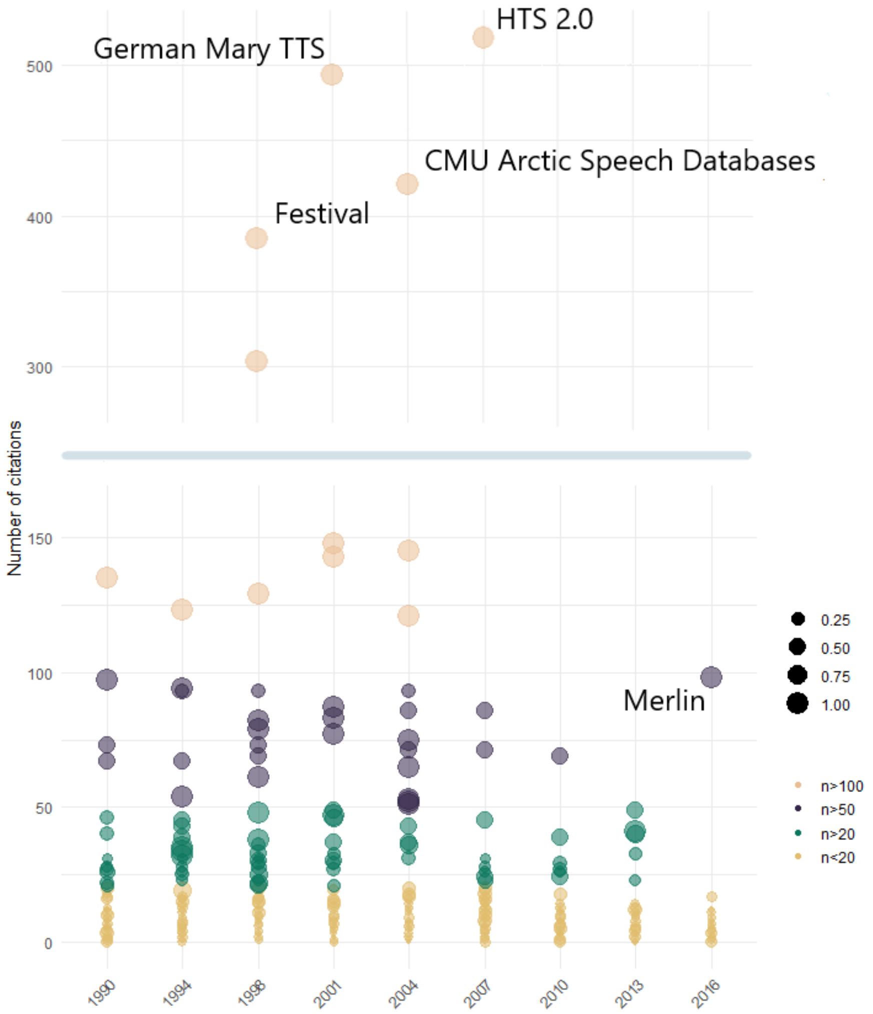

Figure 2 – Number of citations per period (given on x-axis): circle size indicates number of citations, colors correlate with number of citations, names of most-cited papers are given. The blue line indicates a break in the y-axis.

The most prominent period based on the absolute number of citations is the SSW 3 (1998). With the introduction of the US system *Festival Speech Synthesis* [10] on one hand, but also the first approaches for HMM based synthesis [13] on the other, this period dealt with different essential problems in speech synthesis at the same time. Also what differentiates this period from the others is that each paper was cited at least once, whereas other periods include several contributions without any citation.

4.3 Affiliation

Who is behind the progress in speech synthesis? The SSW thought collective is characterized by different fields and aspirations. After looking at the affiliated institutions of the authors, we could identify four main backgrounds. Small companies (not on the stock exchange), research institutes, which are not designated for a sole scholarly mission, universities, and corporations (on the stock exchange). While companies maintain a rather small percentage of participation over the years, as well as research institutions, we can recognize different levels of participation in the categories university and corporations. In 1998 corporations and universities were participating at the same extent, another argument for the importance and versatile interests of the late 1990s speech synthesis research. However, the corporations' participation rapidly drops in the following years. This decrease interestingly coincides with the introduction of open-source license models in the early 2000s, which may be a reason, since it was not attractive anymore for the industry to get involved. While the corporations' interest in contributing diminishes with open-source models, the academic interest grows, reflecting how shared and accessible knowledge is valued highly in a scholarly context. Another important factor is the fluctuation of resources available to individual institutions. Since several institutions were closed over the years, the frequent rise and subsequent decrease of the category research can also be interpreted as a result of irregular funding. Taking a closer look at the individual involvement of the institutions, we can identify three university departments to be the centers of the current speech synthesis research. While the Center of Speech Technology Research (CSTR), University of Edinburgh, was always a prominent actor at the SSW, the Language Technology Institute (LTI), Carnegie Mellon University, and the Nagoya Institute of Technology (NiTech) distinguished themselves in the early 2000s. It has to be noted, that the visibility of certain institutions is often affected by personnel changes and is also strongly linked to individual researchers as well as institutional decisions.

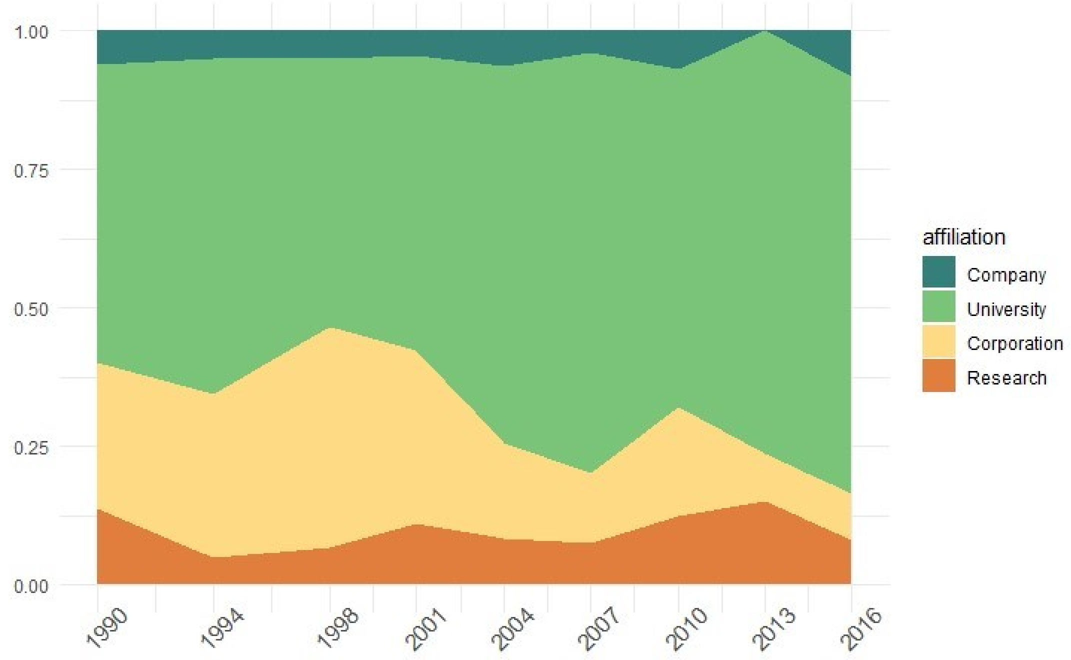

Figure 3 – Affiliation of SSW authors per period

5 Conclusion

With our analysis of key terms, we traced the development of SSW contribution in the course of the last three decades. Unsurprisingly, we found a shift in the number of occurrences a corollary to the advancement of the field, shifting the focus points of the whole community. While personnel and institutions involved are constantly changing, a number of important centers can be identified, contributing over the whole timespan to the SSW. By their way of expressing their thoughts they have a considerable impact on the behaviour of the SSW thought collective as they set methods state-of-the-art, established or introduce new methods and terms to the discourse.

For the case of the terms "neural network" and "unit selection", we could also show that a term is introduced to a discourse community long before it gains prominence. As the associated methods get established and refined, their usage and attribution change in the same way, as it becomes adapted to the thought style. With the field moving on, however, they become more and more historical and obsolete. While the original terms and methods may fade away on the textual level, ideas established through them still shape later developments. We saw that at first "di-/triphone" concatenation was introduced as an explanatory term when the idea of piecing chunks of speech together was still new, until it was replaced by the more general term "unit selection" as the members of the collective would consider this term to be already familiar with the idea of concatenating phones.

For other ideas, they might only flourish as soon as the means of their application are present, as we saw with "neural networks". Here, advancements outside the field of synthesis together with a shift in research questions prepared the ground for a sudden interest after the term seemingly vanished, but not without a redefinition as *deep*, indicating a change in possibilities and conditions. Very similarly, HMMs became the main focus of submissions as concatenative approaches were in need of new ways to solve open questions, dominating the years 2010 and 2013. By 2016, with DNNs, a new focus point emerges.

We also explored paper citations as a dimension for reception. We found a clear lead for papers describing synthesis systems all across the investigated timespan. Also, we observed a rather sharp decline in overall citations from 2007 to 2010. Partly, this might be linked to the fact that until *Merlin* in 2016, no major systems were presented, but we also suspect a link to the decreasing number of contributing corporations. Clearly, visible human-computer dialogue applications indicate the increasing interest of the telecommunications industry and society in general as speech synthesis technology is on the brink of becoming part of our daily life. The research landscape has expanded towards this development in recent years and we might see the SSW adapting to these new fields thus incorporating new ideas into the thought style yet again.

6 Acknowledgments

We thank Pierre Lefebvre for collecting and processing the corpus data as well as preparing it for the analysis.

References

[1] Fleck, L. and L. Schäfer: Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv. Suhrkamp Taschenbücher Wissenschaft. Suhrkamp Verlag GmbH, 1980.

[2] Anthony, L.: AntConc. 2019. URL `http://www.laurenceanthony.net/software`.

[3] Spitzmüller, J. and I. H. O. Warnke: Diskurslinguistik: Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. De Gruyter Studium. Walter de Gruyter GmbH Co.KG, Berlin/Boston, 2011.

[4] Swales, J.: Genre analysis: English in academic and research settings. The Cambridge applied linguistics series. Cambridge Univ. Press, Cambridge, 1993.

[5] Briggs, C. L. and R. Bauman: Genre, intertextuality and social power. Journal of Linguistic Anthropology, 2(2), pp. 131–172, 1992.

[6] Traber, C.: F0 generation with a data base of natural F0 patterns and with a neural network. In 1st ESCA SSW. Autrans, France, 1990.

[7] Lu, H., S. King, and O. Watts: Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. In 8th ISCA SSW. Barcelona, Spain, 2013.

[8] Wu, Z., O. Watts, and S. King: Merlin: An open source neural network speech synthesis system. In 9th ISCA SSW. Sunnyvale, USA, 2016.

[9] Schröder, M. and J. Trouvain: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. In 4th ISCA SSW. 2001.

[10] Taylor, P. A., A. W. Black, and R. Caley: The architecture of the Festival speech synthesis system. In 3rd ESCA SSW. Jenolan Caves, Australia, 1998.

[11] Zen, H., T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W Black, and K. Tokuda: The HMM-based speech synthesis system (HTS) version 2.0. In 6th ISCA SSW. 2007.

[12] Kominek, J. and A. W Black: The CMU ARCTIC speech databases. In 5th ISCA SSW. 2004.

[13] Tamura, M., T. Masuko, K. Tokuda, and T. Kobayashi: Speaker adaptation for Hmm-based speech synthesis system using MLLR. In 3rd ESCA SSW. Jenolan Caves, Australia, 1998.