The Speech Alignment Paradox

David Sündermann[1], *Jaka Smrekar*[2], *Harald Höge*[3], *Antonio Bonafonte*[4], *Hermann Ney*[5]

[1]SpeechCycle, Inc., New York City, USA
[2]University of Ljubljana, Ljubljana, Slovenia
[3]Siemens Corporate Technology, Munich, Germany
[4]Technical University of Catalonia, Barcelona, Spain
[5]RWTH Aachen, Aachen, Germany

david@suendermann.com jaka.smrekar@fmf.uni-lj.si
harald.hoege@siemens.com antonio.bonafonte@upc.edu ney@cs.rwth-aachen.de

Abstract

Applying a recently presented text-independent speech alignment technique based on unit selection to the training of a voice conversion system suggested that the more training data was available, the less speaker-specific information was learned. This paradoxical effect contradicts experience we have from other corpus-based applications as speech recognition or synthesis. There, the performance usually gains with increasing amount of data. In this paper, we investigate this paradox by means of several experiments and derive a mathematical proof for a special case of the speech alignment paradox.

Index Terms: speech processing, speech alignment

1. Introduction

In several speech processing applications (e.g. in speech recognition [1], speaker identification [2], or speech data mining [3, 4]), we have to find a time alignment between speech samples, usually generated by different speakers. Mainly, the texts underlying the compared speech samples is identical, which allows for applying dynamic time warping [1] to the problem. If the underlying text is known, forced alignment [5] can be performed, which may lead to more accurate results.

However, certain applications require the alignment of utterances, which are not parallel. Here, we face the text-independent alignment task. Recently, we presented a technique based on unit selection, which was used for text-independent voice conversion training [6] and later extended to cross-language voice conversion [7].

When compared to text-dependent alignment (dynamic time warping), the achieved speech quality of the voice-converted speech was improved by means of the novel technique, whereas the similarity to the target speaker decreased. Table 1 shows the results of a subjective evaluation reported in [7]. As common metrics, for both overall speech quality and similarity to the target, a mean opinion score [8] on a five-point scale (1 for bad to 5 for excellent) was applied.

As informal listening tests suggested, both effects, the quality boost and the similarity score loss, increased with increasing amount of training data. This paper is to study this paradox focusing on the similarity effect, which can be described by objective criteria, rather than the speech quality, whose objective investigation is still a hard problem [9].

This work has been partially funded by the European Union under the integrated project TC-Star - Technology and Corpora for Speech to Speech Translation - http://www.tc-star.org.

	MOS_Q (quality)	MOS_S (similarity)
text-dependent	3.3	2.4
text-independent	3.5	2.0
source voice	4.7	1.6

Table 1: Results of a subjective evaluation on the application of speech alignment to voice conversion: overall speech quality (MOS_Q) and similarity to the target (MOS_S)

2. Text-Independent Speech Alignment Based on Unit Selection

We consider two arbitrary speech samples to be aligned. At first, they are broken down into frames[1]. Now, the frames are encoded leading to two sequences of feature vectors[2] representing source and target speech, x_1^M and y_1^N. To perform the alignment, from the latter, vectors are to be selected and joined to a sequence $\tilde{y}_1^M$ that optimally corresponds to the source sequence. This is done by taking two criteria into account:

- The distance between source and corresponding target features (*target cost*) is minimum (optimal correspondence).
- The distance to the neighbors of the corresponding target feature vector (*concatenation cost*) is minimum (optimal naturalness). This criterion is to select naturally smooth segments[3] from the target feature vector sequence y_1^M.

Mostly, these optima do not coincide, and we must get by with a compromise between both: We search for the minimum of the weighted sum of target and concatenation cost for each source feature vector:

$$\tilde{y}_1^M = \arg\min_{y_1^M} \sum_{m=1}^{M} \Big\{ \alpha S(y_m - x_m) + (1-\alpha) S(y_{m-1} - y_m) \Big\}. \quad (1)$$

[1]In our study, we utilized pitch-synchronous time frames produced by the Praat tool [10], since this allows for using standard pitch modification techniques to change prosodic properties of speech in the framework of voice conversion. However, all the following considerations also apply to constant frame lengths as mostly used in speech recognition.

[2]Here, we use line spectral frequencies; in other applications, one would certainly prefer other types as mel frequency cepstral coefficients or linear predictive coefficients, cf. [11].

[3]or *units*; that is, where the term *unit selection* comes from. This paradigm is well-known from concatenative speech synthesis, where optimal speech units are selected and concatenated, cf. [12].

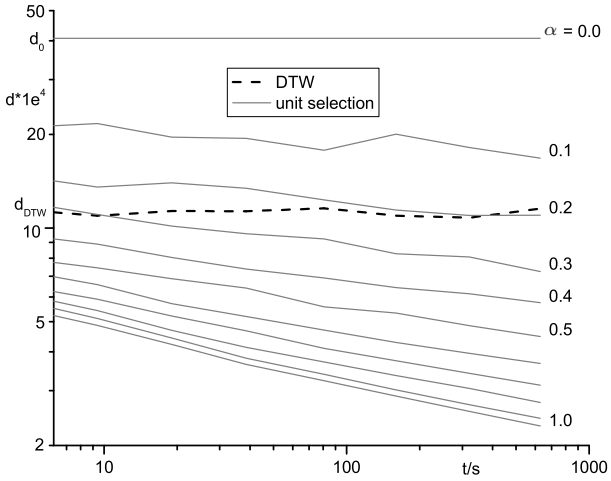

Figure 1: Text-independent speech alignment: average distance between corresponding source and target feature vectors d depending on the amount of data and the trade-off parameter α.

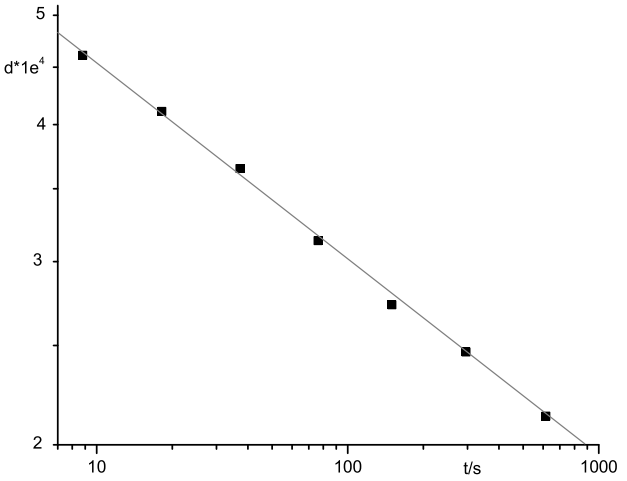

Figure 2: Special case $\alpha = 1$.

Here, $S(y - x)$ is the Euclidean distance between the vectors x and y, and $0 \leq \alpha \leq 1$ is a weight influencing the trade-off between target and concatenation cost.

3. Experimental Evidence of the Speech Alignment Paradox

As already argued in Section 1, we want to limit the investigations on the speech alignment paradox to the similarity of the aligned speech samples. We claimed that the more data was available, the less speaker-specific information could be extracted for the application to voice conversion. An explanation of this effect is that the units, which are selected to minimize the Euclidean distance to the source become more and more similar to the latter, the more data is available to select from. Hence, it provides less and less target-specific information.
To investigate this effect, we want to use the mean Euclidean distance between the aligned feature vector sequences as an objective measure:

$$d = \frac{1}{M} \sum_{m=1}^{M} S(\tilde{y}_m - x_m) \; .$$

Now, we want to look at the dependence of the increasing similarity between source and aligned target speech, i.e. decreasing d value, on the amount of data available. In doing so, we also have to take the trade-off parameter α, see Eq. 1, into account. We conducted experiments using the evaluation corpus of the European speech-to-speech translation project TC-Star [13], which consists of about 10 minutes of speech of two female and two male British English voices. Independent of the voice combinations to be aligned, we got very similar outcomes. As an example, we display the results of a female-male voice combination in Figure 1 in double logarithmic representation. We observe that independent of the trade-off parameter α, the values of d almost constantly decrease[4]. To simplify matters, in the following, we look at the special case $\alpha = 1$; the respective diagram is shown in Figure 2.

[4]except for $\alpha = 0$, which does not lead to a useful alignment, since no target costs are considered

For the considered amounts of data, our test samples are almost located on a straight line in double logarithmic representation. Consequently, the relation between d and t can be approximated by[5]:

$$\log d = c - b \log t \quad \text{with} \quad b > 0 \; ;$$

exponentiation yields

$$d = e^{c - b \log t} = e^{c} e^{\log t^{-b}} = a t^{-b} \quad \text{with} \quad a, b > 0 \; . \quad (2)$$

If we assume the validity of Eq. 2 also for amounts of data beyond the experiment's scope, we get the limit

$$\lim_{t \to \infty} d = \lim_{t \to \infty} a t^{-b} = 0 \; . \quad (3)$$

This means, for very large amounts of data, the aligned speech samples become very similar to each other (for the limit case even identical), which provides evidence for the speech alignment paradox. Unfortunately, the speech alignment algorithm based on unit selection is very computationally expensive (cf. [7]); to process 400 seconds of speech, the computation took more than 80 hours on a 3GHz Intel Xeon machine. Thus, currently, we are not able to massively increase the amount of data involved. This is the main reason for describing the paradox by mathematical means as done in the next section.

4. Towards a Mathematical Proof of the Speech Alignment Paradox

Although the empirical investigations of Section 3 were confirmed by several experimental cycles, doubts arose on the validity of the limit value shown in Eq. 3, as it could be interpreted as follows:
If there is enough speech data available, an arbitrary utterance of an arbitrary voice can be produced only by selecting and concatenating units from this data.
However, the crucial point in the statement is the word *enough*. Applying the parameters $a = 6.8$ and $b = 0.18$ determined on the data of Figure 2 to Eq. 2, we estimated the required amount of data for several degrees of similarity, cf.

[5]in the following equations, we use the normalized time $t := \frac{t}{\mathrm{s}}$ to avoid confusion

d	t	disk space
5	5.6 s	174 kB
2	900 s = 15 min	27 MB
1	$4.2 \cdot 10^4$ s = 11.7 h	1.3 GB
0.5	$2.0 \cdot 10^6$ s = 22.8 d	59 GB
0.2	$3.2 \cdot 10^8$ s = 10.3 a	9.2 TB

Table 2: Required amount of data (t) for certain degrees of similarity (d) and the corresponding hard disk space necessary for storing a 16kHz/16bit PCM version of the data

Table 2. We see that the amount of necessary data extremely grows when the mean distance between source and aligned target feature vectors becomes smaller and soon exceeds the limits of the technical possible.
Nonetheless, since the validity of the statement phrased above could be of high interest to the speech processing community, in the following, we will investigate the alignment technique's behavior for very large amounts of data by mathematical means.

4.1. Speech as a Mixture of Gaussians

As introduced in Section 2, we describe the processed speech by sequences of feature vectors, whose statistical characteristics are very often described by means of the Gaussian mixture model – in literature, we find applications of this model to speech recognition [14], language identification [15], voice conversion [16], speaker recognition [17], speaking rate estimation [18], gender classification [19], etc.
The success of the Gaussian mixture model in these speech processing fields also suggests its application to the investigation of the speech alignment paradox.
In order to keep things manageable, we reduce the number of degrees of freedom as follows:
• We set the number of Gaussian mixture densities to $K = 1$[6].
• We reduce the dimensionality of the feature vectors to $D = 1$ (w.l.o.g.).
• We assume identical covariance matrices for the feature vector sequences to be aligned, i.e., for $D = 1$, we have the standard deviation σ.

4.2. Proving a Special Case

To determine the expected value of the distance d between a source feature vector x and the closest of N target feature vectors y_1^N, we exploit the fact that x is normally distributed with mean μ_x and standard deviation σ and reduce the problem to determining the expected value of d given x:

$$E_N(d) = \int_{-\infty}^{\infty} E_N(d|x)\mathcal{N}(x|\mu_x,\sigma)\mathrm{d}x \ , \tag{4}$$

$\mathcal{N}(x|\mu_x,\sigma)$ is the probability density function of a normal distribution. In the following, we use the *standard* normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and modify Eq. 4 accordingly

$$E_N(d) = \frac{1}{\sigma}\int_{-\infty}^{\infty} E_N(d|x)\, f\Big(\frac{x-\mu_x}{\sigma}\Big)\,\mathrm{d}x \ . \tag{5}$$

Figure 3 shows $E_N(d)$ for several values of N as a function of the normalized distance between the distribution means $\delta = \frac{\mu_y-\mu_x}{\sigma}$.
For the expected value of d if x is fixed, we have

$$E_N(d|x) = \int_{-\infty}^{\infty} |x-y|\, p_N(y|x)\mathrm{d}y \ . \tag{6}$$

Again, we assume the vectors y_1^N to be normally distributed with the parameters μ_y and σ and independent of each other.
For each possible y, we calculate the probability density of the n^{th} target feature vector being equal to y and closest to x. The sum over all of these vectors from 1 to N yields the searched density $p_N(y|x)$.
To be more detailed: The probability density of the n^{th} vector being equal to y is

$$P_n = \frac{1}{\sigma} f\left(\frac{y-\mu_y}{\sigma}\right) \ .$$

The probability of the n^{th} vector being closest to x means that the distance to all other vectors y_ν for $\nu \in \{1,\dots,N\}, \nu \neq n$ is greater than that to y_n or, given $y_n = y$, that $|y_\nu - x| > |y-x|$:

$$\begin{aligned}
Q_n &= p\Big(\bigwedge_{\substack{\nu=1\\ \nu\neq n}}^{N} |y_\nu - x| > |y-x|\Big) \\
&= \prod_{\substack{\nu=1\\ \nu\neq n}}^{N} p(|y_\nu - x| > |y-x|) \\
&= p(|\psi - x| > |y-x|)^{N-1} \\
&= \begin{cases} p(\psi < y \ \vee \ \psi > 2x-y)^{N-1} & \text{for } y < x \\ p(\psi > y \ \vee \ \psi < 2x-y)^{N-1} & \text{otherwise} \end{cases} \\
&= \begin{cases} \Big(\Phi\Big(\frac{y-\mu_y}{\sigma}\Big) + 1 - \Phi\Big(\frac{2x-y-\mu_y}{\sigma}\Big)\Big)^{N-1} & \text{for } y < x \\ \Big(1 - \Phi\Big(\frac{y-\mu_y}{\sigma}\Big) + \Phi\Big(\frac{2x-y-\mu_y}{\sigma}\Big)\Big)^{N-1} & \text{otherwise.} \end{cases}
\end{aligned}$$

Here, ψ is a y-like distributed random variable replacing y_ν for $\nu \in \{1,\dots,N\}, \nu \neq n$; and $\Phi(x)$ is the standard normal cumulative density function, thus we have $\frac{\mathrm{d}\Phi(x)}{\mathrm{d}x} = f(x)$. Accordingly, Eq. 6 becomes

$$\begin{aligned}
E_N(d|x) &= \int_{-\infty}^{\infty} |x-y| \sum_{n=1}^{N} (P_n Q_n)\mathrm{d}y \qquad (7) \\
&= \frac{N}{\sigma}\int_{-\infty}^{x} (x-y) f\Big(\frac{y-\mu_y}{\sigma}\Big)\Big(1 + \Phi\Big(\frac{y-\mu_y}{\sigma}\Big) - \Phi\Big(\frac{2x-y-\mu_y}{\sigma}\Big)\Big)^{N-1}\mathrm{d}y \\
&\quad - \frac{N}{\sigma}\int_{x}^{\infty} (x-y) f\Big(\frac{y-\mu_y}{\sigma}\Big)\Big(1 - \Phi\Big(\frac{y-\mu_y}{\sigma}\Big) + \Phi\Big(\frac{2x-y-\mu_y}{\sigma}\Big)\Big)^{N-1}\mathrm{d}y
\end{aligned}$$

[6] Hence, for these considerations, there is no need for using the term *mixture* when referring to the model. However, the authors showed that the proof can also be derived for arbitrary numbers of Gaussian mixture densities for the source and target speech. To print the full proof would be beyond the scope of this publication, but the authors would be happy to provide it on demand.

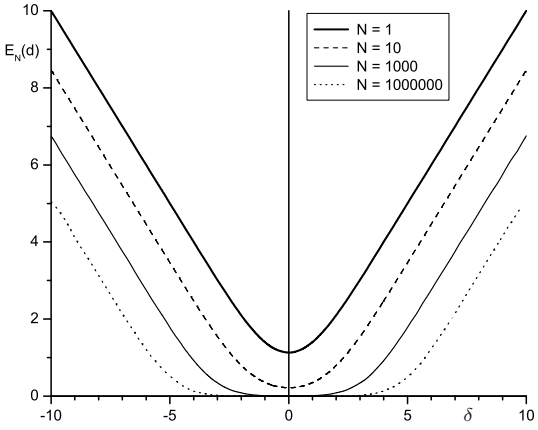

Figure 3: Expected value of the minimum distance between a source feature and N target feature vectors $E_N(d)$ as a function of the normalized difference of the distribution means δ and the number of available target feature vectors N; $\sigma = 1$.

Substituting $u = \frac{y-\mu_y}{\sigma}$ and $v = 2\frac{x-\mu_y}{\sigma} - u$ in Eqs. 5 and 7, we derive

$$E_N(d) =$$

$$\frac{N\sigma}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{v} f\Big(\frac{u+v}{2}+\delta\Big)(v-u)f(u)(1+\Phi(u)-\Phi(v))^{N-1}\mathrm{d}u\,\mathrm{d}v$$

$$-\frac{N\sigma}{4}\int_{-\infty}^{\infty}\int_{v}^{\infty} f\Big(\frac{u+v}{2}+\delta\Big)(v-u)f(u)(1-\Phi(u)+\Phi(v))^{N-1}\mathrm{d}u\,\mathrm{d}v.$$

An evident application of integration by parts in the inner integrals – the terms $Nf(u)(1+\Phi(u)-\Phi(v))^{N-1}$ and $-Nf(u)(1-\Phi(u)+\Phi(v))^{N-1}$ are differentials of $(1+\Phi(u)-\Phi(v))^N$ and $(1-\Phi(u)+\Phi(v))^N$ with respect to u – yields the equality

$$E_N(d) =$$

$$\frac{\sigma}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{v}(1+\Phi(u)-\Phi(v))^N f\Big(\frac{u+v}{2}+\delta\Big)\Big(\Big(\frac{u+v}{2}+\delta\Big)\frac{v-u}{2}+1\Big)\mathrm{d}u\,\mathrm{d}v$$

$$+\frac{\sigma}{4}\int_{-\infty}^{\infty}\int_{v}^{\infty}(1-\Phi(u)+\Phi(v))^N f\Big(\frac{u+v}{2}+\delta\Big)\Big(\Big(\frac{u+v}{2}+\delta\Big)\frac{v-u}{2}+1\Big)\mathrm{d}u\mathrm{d}v.$$

Switching the roles of u and v in the second summand applying Fubini's theorem, we finally obtain

$$E_N(d)=\frac{\sigma}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{v}(1+\Phi(u)-\Phi(v))^N f\Big(\frac{u+v}{2}+\delta\Big)\mathrm{d}u\,\mathrm{d}v.$$

Note that for $u \le v$ the expression $1+\Phi(u)-\Phi(v)$ is at most 1, and therefore

$$\lim_{N\to\infty}(1+\Phi(u)-\Phi(v))^N = \begin{cases}1 & \text{for } u = v\\ 0 & \text{otherwise.}\end{cases}$$

A double application of Lebesgue's dominated convergence theorem yields

$$\lim_{N\to\infty} E_N(d) = 0,$$

which proves the considered special case of the speech alignment paradox.

5. Conclusion

For the highly computational complexity of text-independent speech alignment based on unit selection, we were not able to investigate the speech alignment paradox by means of very large amounts of data. This was the reason for applying a mathematical model describing two speech samples by means of Gaussian mixture models. For a special case, we could derive a mathematical proof of the paradox. Future work is to focus on the generalization of this paper's investigations.

6. References

[1] C. Myers and L. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition," *Bell System Technical Journal*, vol. 60, no. 7, 1981.

[2] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, 1981.

[3] D. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in *Proc. of the AAAI Workshop on Knowledge Discovery in Databases*, Seattle, USA, 1994.

[4] C. Ratanamahatana and E. Keogh, "Everything you Know about Dynamic Time Warping Is Wrong," in *Proc. of the Workshop on Mining Temporal and Sequential Data*, Seattle, USA, 2004.

[5] S. Young, P. Woodland, and W. Byrne, *The HTK Book, Version 1.5*. Cambridge, UK: Cambridge University, 1993.

[6] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-Independent Voice Conversion Based on Unit Selection," in *Proc. of the ICASSP'06*, Toulouse, France, 2006.

[7] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-Independent Cross-Language Voice Conversion," in *Proc. of the Interspeech'06*, Pittsburgh, USA, 2006.

[8] "Methods for Subjective Determination of Transmission Quality," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Recommendation P.800, 1996.

[9] S.-H. Chen, S.-J. Chen, and C.-C. Kuo, "Perceptual Distortion Analysis and Quality Estimation of Prosody-Modified Speech for TD-PSOLA," in *Proc. of the ICASSP'06*, Toulouse, France, 2006.

[10] P. Boersma, "Praat, a System for Doing Phonetics by Computer," *Glot International*, vol. 5, no. 9/10, 2001.

[11] H. Ye and S. Young, "Perceptually Weighted Linear Transformations for Voice Conversion," in *Proc. of the Eurospeech'03*, Geneva, Switzerland, 2003.

[12] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," in *Proc. of the ICASSP'96*, Atlanta, USA, 1996.

[13] H. Höge, "Project Proposal TC-STAR - Make Speech to Speech Translation Real," in *Proc. of the LREC'02*, Las Palmas, Spain, 2002.

[14] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York, USA: Wiley, 1973.

[15] M. Zissman, "Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models," in *Proc. of the ICASSP'93*, Minneapolis, USA, 1993.

[16] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical Methods for Voice Quality Transformation," in *Proc. of the Eurospeech'95*, Madrid, Spain, 1995.

[17] D. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Models," *Lincoln Laboratory Journal*, vol. 8, no. 2, 1995.

[18] R. Faltlhauser, T. Pfau, and G. Ruske, "On-Line Speaking Rate Estimation Using Gaussian Mixture Models," in *Proc. of the ICASSP'00*, Istanbul, Turkey, 2000.

[19] S. Tranter and D. Reynolds, "Speaker Diarisation for Broadcast News," in *Proc. of the Odyssey 2004 Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

[20] A. Kain and M. Macon, "Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction," in *Proc. of the ICASSP'01*, Salt Lake City, USA, 2001.

[21] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A Study on Residual Prediction Techniques for Voice Conversion," in *Proc. of the ICASSP'05*, Philadelphia, USA, 2005.

[22] ——, "A First Step Towards Text-Independent Voice Conversion," in *Proc. of the ICSLP'04*, Jeju Island, South Korea, 2004.